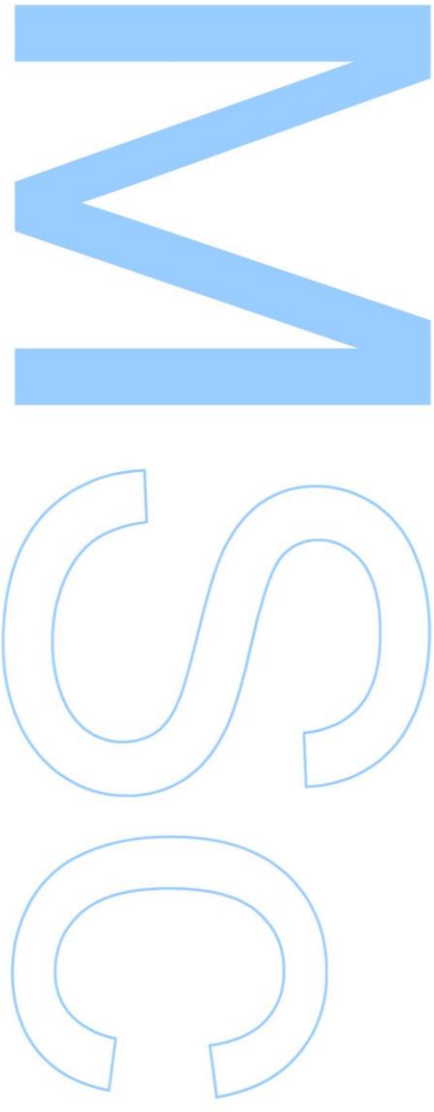
Models and Methods for Motor Insurance Tariffs

Daniela Macedo Correia
Master Degree in Engineering Mathematics
Department of Mathematics
2017

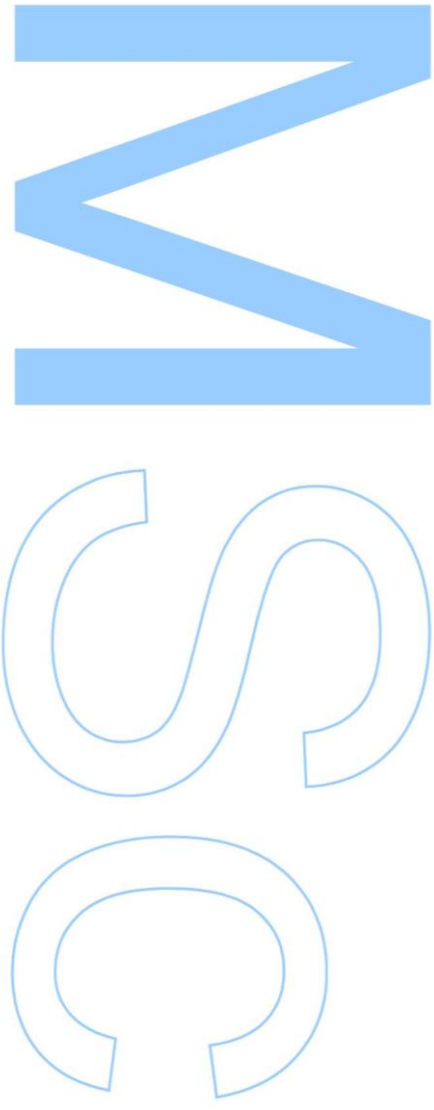
Supervisor
Óscar António Louro Felgueiras
Assistant professor, FCUP



All corrections determined by the jury, and only those, were made.

The President of the Jury,

Porto, _____/_____/_____



To my family

Abstract

In order for an insurance company to grow and prosper it needs to be able to calculate the best estimation of costs for any given policy. As such, an insurance company must estimate as accurately as possible how much it must collect from each policy in order to ensure the coverage of its full costs and also to achieve the desired profit. Moreover, it is clear in every market that no one can afford to lose good customers due to overcharging, therefore being fundamental for a company to create good stable models that predict not only how often clients have claims but also how severe those claims are. The multiplication of these two models, a model for frequency and a model for severity, result in the pure premium, the value that represents the average cost of a policy per year.

In this dissertation, the concept of insurance tariff is presented, as well as all methods and models needed to its creation. We present the classical approach for creating a tariff, the approach nowadays followed in most markets around the world. We also present some ideas and methods for improving this approach, such as methods for dealing with missing data that prevent observations from being excluded, possibly causing bias in the data set. In order to achieve models that better capture the data tendency, Generalized Additive Models (GAM) are also explored, as they allow the introduction of smooth functions.

The last topic of this dissertation consists of a practical approach to apply the techniques explored using the software *R*, R Core Team (2016), and in an attempt to combine them despite the frequent incompatibilities between packages. In that way, a methodology is defined to create the final frequency and severity models. In the end, the classical and the final tariffs are compared, concluding the classical approach to be very reasonable even though both multiple imputation and GAM seem to improve the results. Also, the distributions used in the classical approach are proved to be appropriate, despite the difficulty of achieving models that explain a high percentage of the observed variance using the data set's covariates.

Keywords: Frequency, Severity, Insurance Tariff, Multiple Imputation, Generalized Linear Models, Generalized Additive Models.

Contents

Abstract ii

List of Tables vi

List of Figures viii

1 Introduction 1
- 1.1 The Concept of Insurance . . . 2
- 1.2 Pure Premium . . . 3
  - 1.2.1 Risk Models: The Usual Method . . . 3
- 1.3 Thesis Skeleton . . . 5

2 Data Set Preparation 6
- 2.1 Ideal Structure and Dimension . . . 6
- 2.2 Our Data Set . . . 8

3 Dealing with Missing Values 10
- 3.1 Single Imputation . . . 11
- 3.2 Multiple Imputation . . . 13
  - 3.2.1 Imputation by Chained Equations . . . 13
  - 3.2.2 Analysis and Pooling . . . 14
  - 3.2.3 How Many Imputations Are Needed? . . . 15
- 3.3 Multiple Imputation in Practice . . . 16
  - 3.3.1 Comparison of Variables' Gross Effects Before and After Imputation . . . 23
  - 3.3.2 Conclusions . . . 26

4 Proposed Regression Models 27
- 4.1 Generalized Linear Models . . . 27
  - 4.1.1 Main Results on GLM . . . 30
    - 4.1.1.1 Parameters Estimation . . . 30
    - 4.1.1.2 Deviance . . . 31

4.1.1.3 Residuals . . . 31
4.1.1.4 R^2, Pearson's Pseudo R^2 and Relative Absolute Error 32
4.1.1.5 Hypothesis Testing . . . 33
4.1.1.6 AIC and BIC . . . 34
4.1.2 Overdispersion in Poisson GLM . . . 35
4.2 Zero-Inflated Models . . . 36
4.2.1 Zero-Inflated Poisson Models . . . 38
4.2.2 Zero-Inflated Negative Binomial Models . . . 38
4.3 Generalized Additive Models . . . 39
4.3.1 An Introduction to Smooth Functions . . . 40
4.3.1.1 Splines . . . 40
4.3.1.2 Controlling the Degree of Smoothing with Penalized Regression Splines . . . 42
4.3.1.3 Two More Things About Splines . . . 44
4.3.2 Final Notes About GAM . . . 46

5 Risk Models Using R 47
5.1 Classical Approach . . . 48
5.1.1 Frequency Models . . . 48
5.1.2 Severity Models . . . 49
5.1.3 The Classical Tariff . . . 50
5.2 Potential Problems with Frequency Models . . . 54
5.2.1 Model Comparison . . . 56
5.3 The Ideal Severity Model Distribution . . . 59
5.3.1 Some Conclusions About the Severity Model . . . 64
5.4 Smooth Functions . . . 66
5.4.1 Time Effect . . . 66
5.4.2 Smoothing Continuous Variables . . . 68
5.4.3 Obstacles: Zero-Inflated Models in `mgcv`. . . . 69
5.5 The Final Tariff . . . 70
5.5.1 Methodology . . . 70
5.5.2 Comparison of Tariffs . . . 72

6 Conclusions 75
6.1 Main conclusions . . . 75
6.2 Future Work . . . 77

References 78

List of Tables

3.1 Guidelines according to Graham et al. (2007) for the ideal relation between the number of imputations m and the fraction of missing information λ. . . . 16
3.2 Gross effects of categorical imputed variables on cost of claims, before and after imputation. . . . 25

4.1 Some Exponential Family distributions and their canonical link functions. 29

5.1 Coefficients, standard error and p-values of **Freq M1**. ([1])Fitted variables in decades. ([2])Fitted variable divided by 100. . . . 51
5.2 Coefficients, standard error and p-values of **Sev M1**. ([1])Fitted variables in decades. ([2])Fitted variable divided by 100. . . . 52
5.3 Final tariff relativities for combined models **Freq M1** and **Sev M1**. ([1])Fitted variables in decades. ([2])Fitted variable divided by 100. . . . 53
5.4 Comparison of models **Freq M1**, **Freq M2**, **Freq M3** and **Freq M4** by its mispredictions, Pearson's residuals, AIC and BIC. . . . 56
5.5 Vuong tests between the four candidates for better frequency model - **Freq M1**, **Freq M2**, **Freq M3** and **Freq M4**. . . . 57
5.6 Coefficients, standard error and p-values of **Freq M3**, a ZIP model with two components, a logistic model that generates structural zeros (on the left) and a Poisson count model (on the right). ([1])Fitted variables in decades. ([2])Fitted variable divided by 100. . . . 58
5.7 Cumulated cost of claims and number of claims on the upper tail distribution, by distribution quantiles. . . . 59
5.8 Average quantile absolute error (AQAE) of some fitted distributions to cost of claims data, without large claims. . . . 62
5.9 Average quantile absolute error (AQAE) of Gamma and Normal distributions fitted to log-transformed cost of claims, without large claims. . . . 63
5.10 Summary of observed costs and predicted costs by models **Sev M2** and **Sev M3**. . . . 64
5.11 Comparison of two ZIP models estimated using functions **zeroinfl()** and **ziplss()** with their analog GLM model. . . . 69

5.12 Comparison between GLM and GAM for the final tariff. 73
5.13 Observed Frequency and Severity in test data sets. 74
5.14 Comparison between models with and without multiple imputation. . . 74

List of Figures

2.1 Example of an ideally structured data set. 7

3.1 Flowchart of multiple imputation for $m = 3$ 13
3.2 Barplot of number of articles published in health fields available at *PubMed* containing the words "multiple imputation", by publication year, until August 2017. 16
3.3 Histogram, patterns and percentages of combinations of variables with missing values, given by the **aggr()** function. 17
3.4 Marginplots for some combinations of database's continuous variables. . 18
3.5 Histograms of database's continuous variables before and after transformation. 19
3.6 Density plots of the imputed value (magenta) and the observed values (blue) for all continuous variables with imputed values. 22
3.7 Stripes plots of the imputed value (magenta) and the observed values (blue) by imputation for all continuous variables. 22
3.8 Gross effects of untransformed continuous imputed variables on cost of claims, before and after imputation. 25

4.1 Sketch of the underlying principle of ZI models (Figure inspired in Zuur et al. (2009)). 37

5.1 Data distribution of cost of claims for Own Damage (OD) and Third Part Liability (TPL). 50
5.2 Quantile-plot of severity of claims empirical distribution. 60
5.3 Histogram of cost of claims and some fitted distributions. 61
5.4 Quantile-Quantile plot for comparison empirical cost of claims distribution with some fitted distributions, without large claims. 62
5.5 Quantile-Quantile plot for comparison empirical cost of claims distribution with Gamma and Normal distribution, after log-tranform cost of claims, without large claims. 63
5.6 Model output plots - models **Sev M2** (the four plots on the left) and **Sev M3** (four plots on the right). 65

5.7 Predicted cost of claims by models **Sev M2** and **Sev M3** against observed costs. 65
5.8 Gross Effect of variables `Year` and `Month` in severity of claims. 67
5.9 Gross Effect of variable `Year` in frequency of claims. 68
5.10 Tendency of variables `AgeVehicle`, `AgeDriver` and `HP` captured with smooth splines, for the frequency model (top) and the severity model (bottom). 68
5.11 Procedure flowchart. 71
5.12 Final tariff's analysis diagram. 72

Chapter 1

Introduction

As important as it is to modern society, insurance is not a recent idea. The origins of insurance are related to marine transportation and wealthy merchants. It is thought that the earliest form of insurance occurred more than two millennia ago in the Chinese civilization, where merchants along the Yangtze river decided to split shipments into smaller portions and placed them on several boats in order to reduce the risk of losing all cargo at once.

The more formalized insurance arrangements we are familiar with today actually began in the late 1600s, in London, at a coffeehouse owned by Edward Lloyd where merchants gathered. Concerned that they could be devastated financially if an entire shipment was lost, merchants began to make arrangements with each other to share their risks of loss, receiving a bonus if the voyage was successful and paying for if the vessel were lost. This arrangement was the beginning of Lloyd's of London, an institution that has continued to operate in such way for more than 400 years and in which all insurance companies around the world are based.

Nowadays, in most countries, insurance companies are split into Life and Non-life companies. Life companies sell life insurances, long-term investments that pay out a sum of money on the death of the insured person. Naturally, Non-life insurance companies sell insurances that are not determined to be life insurances, such as motor insurance, household, personal accidents, health and general third part liability. In this case, risks are covered for a small period, usually a year. This kind of insurance companies is also called property and casualty insurance, in the United States of America and Canada, and general insurance, in the United Kingdom.

Both Life and Non-life business lines require painstaking analysis in order to measure and manage risks and uncertainty. Pricing and reserving departments are set up to for that purpose. The reserving department has the mission of calculating the fund reserve, i.e. the amount needed to be kept aside to cover extreme losses, therefore

avoiding bankruptcy. As for the pricing department, it is responsible for defining how much the company must charge in advanced to costumers. Thus, it is in this department that the insurance tariffs are framed.

1.1 The Concept of Insurance

It all starts when a large number of people face a similar economic risk with a reasonably low probability of occurrence. This risk can be of such magnitude or severity that, even if improbable, its occurrence might dramatically impair the wealth of those who see it materialized, leading to risk aversion and thus a desire to avoid it. The nature of the risk may be diverse. For example, the driver of a car must accept the personal risk that he may be injured in the event of an accident and the financial risk that he may damage the car to the extent that it loses its economic value. The difference between these two risks is that, being the latter financial, it can be transferred. The transfer of financial consequences is the first step in the insurance concept.

The financial consequences of a risk are thus transferred from a number of individuals to a collective fund. This fund includes the collective risk of its members, together with the resources that those members have made available to cope with the occurrence of such risks. So, each member contributes a small amount to the fund to attend to collective loss. The term *common fund* refers to the fact that the risks involved, although not necessarily identical, are of a similar nature. So when we buy an insurance policy we are actually *paying for everybody's mistakes* and the reason we do it willingly is not knowing when we will need to resort to the fund. Faced with the possibility of serious financial loss, it makes sense to opt for a small but certain loss.

Even though it is true that we would all like to be financially protected by a common fund, we would also like to contribute to this fund with the minimum possible amount, which could result in insufficient collective contributions to cover collective losses. Hence, the fee to be charged to each member should be carefully considered. A first approach would be to establish a fixed fee equal to the average cost of loss. However, since the risks are similar but not identical, this method would benefit bad risks and encourage the good risks to leave the fund since the below-average risks are bankrolling the above-average risks, causing the average cost of injury to increase, which may lead to a lack of funding from the fund. An alternative is to charge each member of the fund an amount representing the degree of severity of the risk transferred, i.e, an equitable premium. This requires analysis so that frequency and likely severity of the loss can be estimated for each risk. The expected cost of loss is its likely frequency multiplied by its likely severity. An amount equivalent to that expected cost may then

be deposited in the fund, known as the pure risk premium. Using this method, each member of the fund will pay a fair premium based on the transferred risk, promoting balance and avoiding selection of fund members.

This is exactly what happens in markets and what insurers do. It is very rare for an insurer to apply a fixed rate to all clients of a particular portfolio, and when that happens it is because that portfolio is extraordinarily homogeneous. Nowadays, however, given the massification of databases and computerized quotation systems, insurance companies are able to classify and distinguish risks at a minimum cost and as a result equitable premium makes the rule.

1.2 Pure Premium

In order for an insurance company to grow and prosper it needs to be able to calculate the best estimation of costs for any given policy, e.g. losses, expenses, cost of capital and profit loading. In other words, each company must estimate as accurately as possible the premium that each policy should pay to cover its full cost and to get the target profit, known as technical price (TP).

Technical price is able to lead management towards significant opportunities for profitable growth by assessing how big the difference from the paid premium is. This concept, called tariff leakage, is calculated using a best practice risk model.

Risk Models classify, differentiate and calculate pure premium for each risk - the minimum premium necessary to cover the expected losses from a specific risk. Therefore, pure premium is the average cost per policy year or another time period, i.e. $\boldsymbol{pure\ premium = claim\ frequency \times claim\ severity}$.

The idea behind risk models is to classify risk and assign it to a certain risk group based on known information such as age group, vehicle power or geographical zone. If so, it is possible to predict future losses based on historical data of those clients who presented similar risks. Even more, it is clear in every market that no one can afford to lose good customers due to overcharging. For this reason, it is fundamental for the company to be confident in the overall tariff level and to identify and address the main sources of tariff leakage.

1.2.1 Risk Models: The Usual Method

Motor insurances are sold all over the world and are usually one of the most important portfolios of insurance companies. Motor insurances are sold by coverages, i.e. sets of hypothetical situations for which the insurer is providing protection. Examples

of different coverages are: Third Part Liability (TPL), which covers the cost of any property damaged as well as medical bills from resulting injuries to third parties in case of an accident caused by the insured person; Own Damage (OD), which covers the repairs of the insured car; Theft, Fire, Vandalism, Travel Assistance, as well as lots of other coverages. So, if the company A sells two coverages for motor insurance - TPL and OD - each coverage must be treated separately, which means that a pure premium will be estimated independently and the total pure premium will be the sum of TPL and OD pure premiums.

In Portugal, TPL in motor insurance is mandatory which means that usually the TPL portfolio is greater than other coverages such as OD. For that reason, the risk models for TPL may be more sophisticated. Usually, if the data is enough, more models are made dividing the claims in responsible/non-responsible and even in claims with/without body injuries. This means that, if possible, TPL pure premium may have three components - Material Damage Responsible, Material Damage Non-Responsible and Body Injury. The reason why the responsible/non-responsible question is usually only asked for material damage is that claims with body injuries are much rarer than those with only material damage. Sometimes large claims are also modeled separately due to claim severity heavy-tailed distribution. This further analysis will not be explored here.

Nowadays, Non-life insurance pricing resorts to generalized linear models to create models for frequency and severity. These models are based on a set of basic assumptions:
- Assumption 1 (Policy independence)**.** Consider n different policies and let X_i denote the frequency/severity of policy i. Then the $X_1, ..., X_n$ are independent.
- Assumption 2 (Time independence)**.** Consider n disjoint time intervals and let X_i denote the frequency/severity in time interval i. Then the $X_1, ..., X_n$ are independent.
- Assumption 3 (Homogeneity)**.** Consider any two policies with equal risk profile with the same exposure and let X_i denote the frequency/severity of policy i. Then X_1 and X_2 have the same probability distribution.

It is obvious that these assumptions are debatable: in motor insurance, the possibility of a collision between two cars that are insured by the same company violates the policy independence assumption; homogeneity implies that the only thing that matters is the duration of a policy, not when it starts or ends, which may appear unrealistic since many claim types are subject to seasonal variation. However, the most fallacious assumption seems to be time independence, especially for claim frequency. It is normal that a car driver who has had an accident may drive more carefully in the future or even think twice before reporting another claim since it would mean a heavy aggravation of the premium. Thus, the occurrence of a claim might imply a lower claim frequency in the future.

1.3 Thesis Skeleton

This dissertation is organized as follows. Chapter 2 describes the ideal data set structure and the key variables to construct a tariff. In this chapter, the data set used in this dissertation is also described.

In chapter 3 alternatives to deal with missing data apart from the complete cases analysis are presented. Here, two techniques are developed - single imputation and multiple imputation. We study their advantages and disadvantages and present the results of applying multiple imputation to our data set.

Chapter 4 includes the theoretical basis for generalized linear models, which supports most of the methodologies addressed in this thesis. It is also in this chapter that we present zero-inflated models and generalized additive models.

Chapter 5 describes the construction of a tariff, step-by-step. We start by constructing the tariff according to the classical approach. Then, other models and methods are considered in order to improve the original result, such as correcting data for overdispersion, in the frequency model, and testing the more adequate distributions, in the severity case. Finally, we construct our final tariff based on a developed methodology and compare the final and classical tariffs.

Finally, in Chapter 6 we briefly describe the main conclusions and contributions of this dissertation.

Chapter 2

Data Set Preparation

Insurance pricing has changed drastically in the last decades as the amount of available data increases. A few decades ago, insurance tariffs were made based on underwriters' experience, with little or no statistical analysis at all. Nowadays, it is imperative to record, store and analyze, not only for insurance companies but also for all types of business.

Data is definitely the key to establishing a tariff. In order to reach an excellent pure premium estimation, it is important to have good data and it is necessary to control every aspect of it. For this reason, analyzing, cleaning and understanding the data is often one of the most complicated and time-consuming tasks.

In this chapter, we present the ideal data set structure and the "must-have" variables necessary to construct a tariff. Then, the data set used in the analyzes to come is presented.

2.1 Ideal Structure and Dimension

The required number of observations in the data set must be determined based on the frequency of claims and the volatility of the losses. For example, the period of observation to analyze property damage caused by natural events such as floods, storms, or earthquakes should be long enough to have sufficient claims history, which might mean several decades. However, in our case, the high frequencies and stable costs associated with motor policies shorten the time horizon, making it possible to construct a tariff only with 5 to 10 years of historical data.

The construction of the data set must be made according to the exposure view, where exposure is the duration, in years, for which a policy or cover is active in a certain

	A	B	C	D	E	F	G	H	I	J	K
1	ID	Year	Day From	Day Until	Exposure	Claim indicator	Claim month	Claim cost	variable 1	variable 2	variable 3
2	100001	1980	206	365	0,4356	0	0	0	a1	b3	c1
3	100001	1981	1	34	0,0904	1	2	649,30 €	a1	b3	c1
4	100001	1981	35	205	0,4658	0	0	0	a2	b3	c1
5	100001	1981	206	329	0,3370	1	11	6584,98 €	a2	b3	c1
6	100001	1981	330	365	0,0959	0	0	0	a2	b3	c1
7	100001	1981	1	205	0,5589	0	0	0	a2	b3	c1
8	100002	1980	50	365	0,8630	0	0	0	a3	b1	c2
9	100002	1981	1	49	0,1315	0	0	0	a3	b4	c2
10	100003	1985	113	298	0,5068	0	0	0	a6	b1	c3
11	100004	1984	225	365	0,3836	0	0	0	a1	b1	c5
12	100004	1983	1	88	0,2384	1	3	801,27 €	a1	b1	c5
13	100004	1983	89	147	0,1589	0	0	0	a1	b1	c5
14	100004	1983	148	224	0,2082	0	0	0	a4	b1	c5
15	100004	1983	225	365	0,3836	0	0	0	a4	b1	c5
16	100004	1984	1	224	0,6110	0	0	0	a4	b1	c5
17	100004	1984	225	365	0,3836	0	0	0	a4	b1	c5
18	100004	1985	1	224	0,6110	0	0	0	a4	b1	c5

Figure 2.1: Example of an ideally structured data set.

year. Therefore, all policies with at least one day of exposure in the considered period need to be considered in such a way that each row represents the exposure of the policy within each year. For example, a policy which starts on the first of April of a certain year and has one year of exposure must have one row with 9 months exposure for that same year and another row which has three months exposure for the following year. If for some reason the policy's characteristics change at some point in the middle of a year, a new observation must be created, splitting the previous exposure in two, based on the date of the change. The same is applied if a claim occurs, based on the date of occurrence of the claim.

Once the policy has been divided into the right number of rows, each with the correct exposure, it is possible to start considering the other factors which are related to the policy itself. Usually, a data set for tariff purposes contains 4 types of variables:

- **contract data**, such as premiums, installments, coverages or deductibles;
- **customer data**, such as the age of the driver, years of driving license, claims history, vehicle use or occupation;
- **risk data**, concerning in this case characteristics of the vehicle such as brand, model, age and horsepower;
- **territorial data**, such as postcode, county or district.

Besides that, it is mandatory to include in the data set some other variables, such as policy identification, date of beginning and end of exposure (in order to calculate the associated exposure), the dichotomous variable indicating if there was a claim in the exposure period and, in case there was, the respective month and cost.

Once the data set is created as explained, it is easy to obtain data sets for frequency and severity. The data set for frequency uses all observations and almost all variables,

being only possible to exclude the variables that characterize the claim - its cost and month of occurrence. As for the severity data set, only the observations with claims are used and it is not necessary to include the variable of exposure.

2.2 Our Data Set

The data set used in this dissertation was provided by a Portuguese insurance company that preferred to remain anonymous. This data set, although about a real market and real policies, is only a subset of the original base and only contains information about a specific coverage. In addition, all numerical values have been transformed. Thus, it is important to note that all values obtained through analyses are, nevertheless, fictitious.

The structure of our data corresponds to the structure described in the previous section. The data provided contained about 200,000 observations, all the response variables needed to create a tariff and some other unidentified covariates. After working these data and trying to associate the variables available to variables usually included in a tariff, we obtained the following set of variables:

- `ID`: policy identification;
- `RY`: exposure, in years;
- `Year`: year of exposure;
- `AgeDriver`: quantitative variable, integer;
- `AgeVehicle`: quantitative variable, integer;
- `YearsDLicense`: quantitative variable, integer;
- `HP`: quantitative variable, continuous and positive;
- `Weight`: quantitative variable, continuous and positive;
- `ClassAuto`: qualitative variable, nominal dichotomous ;
- `Brand`: qualitative variable, nominal;
- `Seats`: qualitative variable, ordinal;
- `Use`: qualitative variable, nominal dichotomous ;
- `Fuel`: qualitative variable, nominal dichotomous ;
- `ClaimIndicator`: qualitative variable, dichotomous ;
- `ClaimCost`: quantitative variable, continuous and positive;

- `ClaimMonth`: month of occurrence of the claim;

In the previous list, variables four to ten (`AgeDriver` - `Fuel`) are the variables which were originally unidentified. The names given to them were based on their distribution and based on experience. However, this task can be quite difficult, e.g. if no additional information is provided by the levels' labels, it is very difficult to distinguish between two dichotomous variables. Therefore, it is important to keep in mind that these variables are only used as examples.

The data was randomly divided into training and test data sets, in the proportion 70/30, according to the holdout method (Reitermanov (2010)). Therefore, all models considered in the future will be fitted using the training data set and the test data set will be used at the end of this dissertation to compare and evaluate the final tariff.

Chapter 3

Dealing with Missing Values

Missing data are a common problem in insurance data sets. There are several reasons for missing values, as for example costumers who do not want to disclose data or offices that do not submit the relevant data into their data sets. Nevertheless, its cause is always the same - in order to buy a policy the costumer does not have to provide all the information that he is asked for.

Because missing data affect the quality of analyses as well as the robustness and accuracy of estimates, the most common approach is to exclude those observations, usually referred as **complete case analysis**. However, complete case analyses can lead to biased conclusions. For example if a certain agent has a tendency to fill in only the strictly necessary information to submit a policy, all the policies from that agent will be excluded from the analyses.

According to Sterne et al. (2009), there are three types of missing values:

- **Missing completely at random (MCAR):** There are no systematic differences between the missing values and the observed values.
- **Missing at random (MAR)**: Any systematic difference between the missing values and the observed values can be explained by differences in observed data.
- **Missing not at random (MNAR)**: Even after the observed data are taken into account, systematic differences remain between the missing values and the observed values, i.e. the missing value is related to the reason it is missing.

For example, MAR is when the insured person's occupation is missing because the agent did not ask it whereas MNAR is when the occupation is missing because the insured person is unemployed and for that reason did not want to respond.

3.1 Single Imputation

Imputation is the process of replacing missing data with substituted values and it is a method for dealing with missing values. The major attractiveness of this method is being able to use complete-data methods of analysis on the filled-in data set, instead of using only the complete cases. There is more than one way to impute values, but the most common is single imputation.

Single imputation is a very simple method. Basically, it is replacing missing data of a variable Y with its best-prediction. This prediction depends not only on the data set but also on the distribution of Y and on the analyst's experience. For example, if there are no covariates, the best-prediction of Y may be the observed mean or median, depending on the distribution of the variable. If on the other hand there are covariates, the best-prediction may be a regression model as well as any other prediction model.

Once the imputation is completed, the imputed cells are treated as known, true observed values. Thus, inference based on complete-data methods is now possible to perform without loss of observations. The problem, however, is that single imputation compromises the variability of the data, as shown in the following examples:

Example 3.1.1. (Rubin (1987)) Let Y be a random variable of mean $\bar{Y}$ and y a random sample of Y of size n. More over, let $\bar{y}$ be the random variable that represents the mean of y. Thompson (2012) shows that

$$E(\bar{y} - \bar{Y}) = E(\bar{y}) - E(\bar{Y}) = 0 \quad \text{and} \tag{3.1}$$

$$\text{Var}(\bar{y} - \bar{Y}) = \text{Var}(\bar{y}) = s^2 \left(\frac{1}{n} - \frac{1}{N} \right), \tag{3.2}$$

where s^2 is the sample variance and N the population size. Thus, standard complete-data inference considers that

$$(\bar{y} - \bar{Y}) \sim N\left(0,\ s^2 \left(\frac{1}{n} - \frac{1}{N} \right)\right). \tag{3.3}$$

Suppose now that only n_1 of the n values of Y are actually observed and that we have $n_2 = n - n_1$ missing values, where $\bar{y_1}$ and ${s_1}^2$ are the sample mean and variance of the n_1 observed values. Then, by complete-cases analysis, we have that

$$(\bar{y_1} - \bar{Y}) \sim N\left(0,\ s_1^2 \left(\frac{1}{n_1} - \frac{1}{N} \right)\right). \tag{3.4}$$

Now, suppose that instead of using 3.4, the best-prediction of the missing values of Y are imputed and 3.3 is applied, creating a complete data set where there is no

distinction between observed and imputed values. Under the assumptions being made, the best-prediction of each missing value of Y is $\bar{y_1}$. Therefore, the mean of all n values is $\bar{y_1}$ and the sample variance is ${s_1}^2 \times \frac{n_1-1}{n-1}$, so that

$$(\bar{y_1} - \bar{Y}) \sim N\left(0, s_1^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{n_1 - 1}{n-1}\right). \tag{3.5}$$

The ratio between the variance of 3.5 and 3.4 leads us to the conclusion that the variance of 3.5 is smaller than that of 3.4 by essentially a factor of $\left(\frac{n_1}{n}\right)^2$ for large n_1 and $\frac{N}{n_1}$. Thus, interval estimates of $\bar{Y}$ will be too short leading to potentially severe undercoverage, and test statistics will be too large leading to excessively large significance levels.

In conclusion, single imputation might not necessarily lead to correct centered inferences because it not only assumes the sample size to be n when it is really n_1 but also because it underestimates the variance by a factor of $\frac{n_1}{n}$. □

Example 3.1.2. Following the previous example, imagine now that Y has one or more covariates and that we want to apply single imputation with a regression model. The squared standard error of the model, SE, is given by

$$\mathrm{SE}^2 = \frac{\sum_n (y - \hat{y})^2}{n} \tag{3.6}$$

where y is the observed value, $\hat{y}$ is the estimated value and n the number of observed values.

As in the previous example, imagine now that only n_1 of those n values are observed. By complete-cases analysis, the squared standard error of the regression is

$$\mathrm{SE}^{2^*} = \frac{\sum_{n_1} (y - \hat{y_1})^2}{n_1} \tag{3.7}$$

where $\hat{y}_1$ is the estimated value for the n_1 observed values. Suppose now that the model is used to impute the missing values of Y and that 3.6 is applied. In this case, because imputed data fit perfectly along the regression line, we have that

$$\sum_n (y - \hat{y})^2 = \sum_n (y - \hat{y_1})^2 = \sum_{n_1} (y - \hat{y_1})^2. \tag{3.8}$$

Therefore, because

$$\mathrm{SE}^2 = \frac{\sum_n (y - \hat{y})^2}{n} = \frac{\sum_{n_1} (y - \hat{y_1})^2}{n} < \frac{\sum_{n_1} (y - \hat{y_1})^2}{n_1}, \tag{3.9}$$

we have that

$$\mathrm{SE}^2 < \mathrm{SE}^{2^*}, \tag{3.10}$$

which means that the standard error of a regression model is smaller after single imputation than it was before, revealing loss of data variability. □

In conclusion, single imputation does not take into account uncertainty in imputations. Therefore, it can not reflect the sampling variability about the actual value. Once the missing values are imputed, they are treated as known true observed values, which do not represent the truth. For this reason, inferences based on the imputed data set will be too sharp since the extra variability due to unknown missing values is not being considered.

3.2 Multiple Imputation

Multiple imputation (MI) is a statistical technique that proceeds in three main steps: imputation, analysis and pooling. The idea is to create not one but m imputed data sets, analyze them separately and pool the results into a final result. The imputation procedure can be applied using several techniques; we will focus on chained equations.

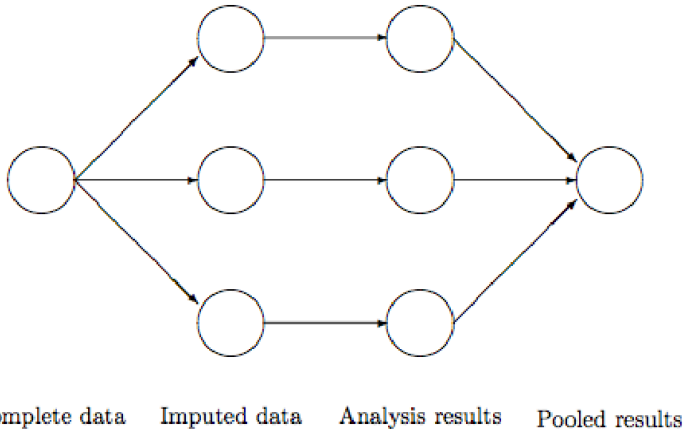


Figure 3.1: Flowchart of multiple imputation for $m = 3$

The original work on multiple imputation can be found in Rubin (1987) and provides excellent insight into many issues in multiple imputation.

3.2.1 Imputation by Chained Equations

To make the approach by chained equations more concrete, imagine a simple example where we have three variables in our data set: X, Y and Z. These variables may or may not have missing values. Suppose, without loss of generality, that all variables have missing values.

Multiple Imputation by Chained Equations (MICE) can be resumed to a simple algorithm. Even though some authors present it as a six-step algorithm (see Azur et al. (2011)), here we present an explanation as a four-step algorithm:

- **Step 1**: For each variable, a single imputation, such as imputing the mean, is performed for every missing value, creating a complete data set;
- **Step 2**: The values imputed for variable X are set back to missing and a regression for X is performed in order to estimate its missing values. In this regression model X is the dependent variable and Y and Z are the complete independent variables, without missing values;
- **Step 3**: Step 2 is repeated for all variables, creating a data set only with observed and regressed values. The cycling of regressing each of the variables consists of one cycle;
- **Step 4**: Steps 2-3 are repeated for a defined number of cycles, with the imputations being updated at each cycle.

These regression models operate under the same assumptions that one would make when performing linear, logistic or Poisson regression. The advantage of this method is having the possibility of adjusting the regression method to each variable, and therefore being able to estimate all different types of variables.

At the end of step 4, the final imputations are retained resulting in an imputed data set. The complete algorithm is then repeated m times, creating m imputed data sets where the originally missing values differ.

3.2.2 Analysis and Pooling

After creating m different data sets, the idea is to analyze them separately and combine the m results into a final result. For example, imagine that the main purpose is to apply a regression model and estimate its parameters. Thus, with multiple imputation, a different regression will be fitted to each one of the m different data sets and m different parameters will be obtained. Then, the point estimate of a certain parameter, say the regression coefficient b, is simply the average of the parameter estimate obtained over the m data sets,

$$b = \frac{\sum_{k=1}^{m} b_k}{m}, \tag{3.11}$$

where b_k is the regression coefficient for the k-th data set. Variance, however, is partitioned into the within imputation variance U_b, which captures the usual sampling variability, and the between imputation variance B_b, which captures the estimation

variability due to missing data. Graham et al. (2007) define these measures as

$$U_b = \frac{\sum_{k=1}^{m} SE_{b_k}^2}{m} \quad \text{and} \quad B_b = \frac{1}{m-1}\sum_{k=1}^{m}(b_k - b)^2, \tag{3.12}$$

where $SE_{b_k}^2$ is the squared standard error for the regression coefficient b in the k-th data set. The combined variance is then given by

$$T_b = U_b + \left(1 + \frac{1}{m}\right) \times B_b \tag{3.13}$$

and the standard error associated to the coefficient b is $SE_b = \sqrt{T_b}$.

The parameter estimate is then divided by its SE to give a t-value that, along with its degrees of freedom, defined as

$$df = (m-1)\left(1 + \frac{m \times U_b}{m+1} B_b\right), \tag{3.14}$$

may be used for statistical inference.

3.2.3 How Many Imputations Are Needed?

The fraction of missing information, λ, is another very important quantity in multiple imputation. Schafer and Olsen (1998) define it as

$$\lambda = \frac{r + \frac{2}{df+3}}{r+1}, \quad \text{where} \quad r = \frac{(1 + \frac{1}{m}) \times B_b}{U_b}.$$

This ratio is especially important to measure the efficiency of an estimation. Rubin (1987) showed that the relative efficiency of an estimate based on m imputations to one based on an infinite number of them is approximately $(1 + \lambda/m)^{-1}$. With 50% missing information, an estimate based on $m = 5$ imputations has a standard deviation that is only about 5% wider than one based on $m = \infty$, since $\sqrt{1 + 0.5/5} = 1.049$. Basically, the author defends that, unless rates of missing information are unusually high, there are no practical benefits of using more than five to ten imputations.

On the other hand, Graham et al. (2007) defends that users of multiple imputation should ask for many more imputations than what has previously thought to be needed. The authors claim that it depends not only on λ but also on one's tolerance for what they say to be the preventable power falloff due to choosing m too small. The authors state that if one is willing to tolerate a 5% power falloff, then the number of imputed data sets should vary from $m = 3$ to $m = 40$, depending on λ. If otherwise one is willing to tolerate only 1% power falloff, then one should use at least $m = 20$, possibly reaching 100 imputations for high values of λ. In this dissertation, we will follow the guidelines defined by Graham et al. (2007), presented in table 3.1.

5% power fallout		1% power fallout	
m	λ	m	λ
3	0.1	20	0.1
10	0.3	20	0.3
10	0.5	40	0.5
20	0.7	40	0.7
40	0.9	100	0.9

Table 3.1: Guidelines according to Graham et al. (2007) for the ideal relation between the number of imputations m and the fraction of missing information λ.

3.3 Multiple Imputation in Practice

MI is a technique used and defended by several data analysts nowadays. Throughout the years, this method gained more and more supporters and is now the principal technique used in several fields to deal with missing values. In fact, multiple imputation has become the standard when dealing with missing data in fields such as ecology, epidemiology and public health, and some reputable journals such as *Nature*, *The Lancet* and *The British Medical Journal* have several articles with this method.

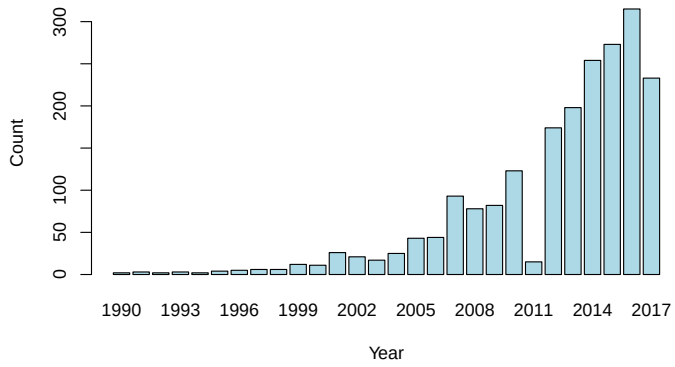


Figure 3.2: Barplot of number of articles published in health fields available at *PubMed* containing the words "multiple imputation", by publication year, until August 2017.

On the other hand, multiple imputation is not a very common method when the subject matter is insurance data sets, even though there are no objections to its use. Since actuarial science already assumes that missing values in insurance data sets are of type MCAR or MNAR, there is nothing to prevent us from using multiple imputation, even though there are a few "rules" to determine whether its use is appropriate or not to a certain data set.

So, the aim of this chapter is to test if multiple imputation, in particular the MICE

method, is correctly filling the gaps in our data set. To do it so, we will compare the gross effect of some variables before and after multiple imputation.

The implementation of the MICE algorithm may be done using the package `mice`, van Buuren and Groothuis-Oudshoorn (2011), of the software *R*, R Core Team (2016). This package provides lots of different tools to analyze the adequacy of the method and to implement it. First of all, it is important to understand the distribution of missing values in our data set. For that, the function **md.pattern()** may be useful, since it returns, as the name suggests, the pattern of missing values. Basically, it summarizes which and how many combinations of variables have missing values.

Another useful and more visual tool is the **aggr()** function. This function provides the same information of **md.pattern()** but in a graphical way, which makes it easier to interpret. The output of this function is shown in figure 3.3. On the left side, the figure displays a histogram with the amount of missing values per variable. In the presented case, it is possible to see that variables `YearsDLicense` and `AgeDriver` are those with more missing values, around 40%. On the right side of the figure, the patterns of missing values (combination of variables with missing values) are presented, in green, together with their percentage. Analyzing this graphic, we can conclude that almost 60% of observations are not missing any information and that 37% have missing values exactly in the variables `YearsDLicense` and `AgeDriver`.

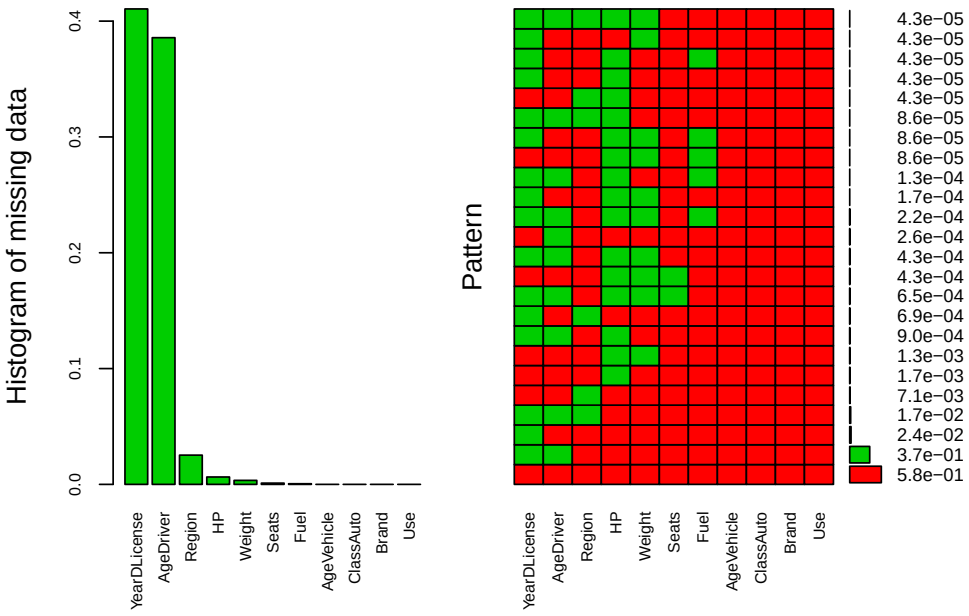


Figure 3.3: Histogram, patterns and percentages of combinations of variables with missing values, given by the **aggr()** function.

Another helpful visual approach is a special plot provided by the function **margin-**

plot(). Here, we are constrained at plotting only two variables at a time, but nevertheless we can gather some interesting insights. This function is especially useful to analyze variables that seem to have related missing values, as it is the case of the variables `YearsDLicense` and `AgeDriver`. Figure 3.4 consists on four marginplots. For example, in the first plot, on the left side, the red boxplot shows the distribution of `YearsDLicense` with `AgeDriver` missing, while the blue boxplot shows the distribution of the remaining data points. Also, the number in the lower left corner, in dark red, is the number of observations that are missing in both variables, being the numbers above and at right, in red, the number of missing values in each variable. It is important to notice that, if the data are assumed to be MCAR, then blue and red boxplots are expected to be very similar. If otherwise MAR is assumed, differences in these boxplots might say that missing values in one variable are explained by the other.

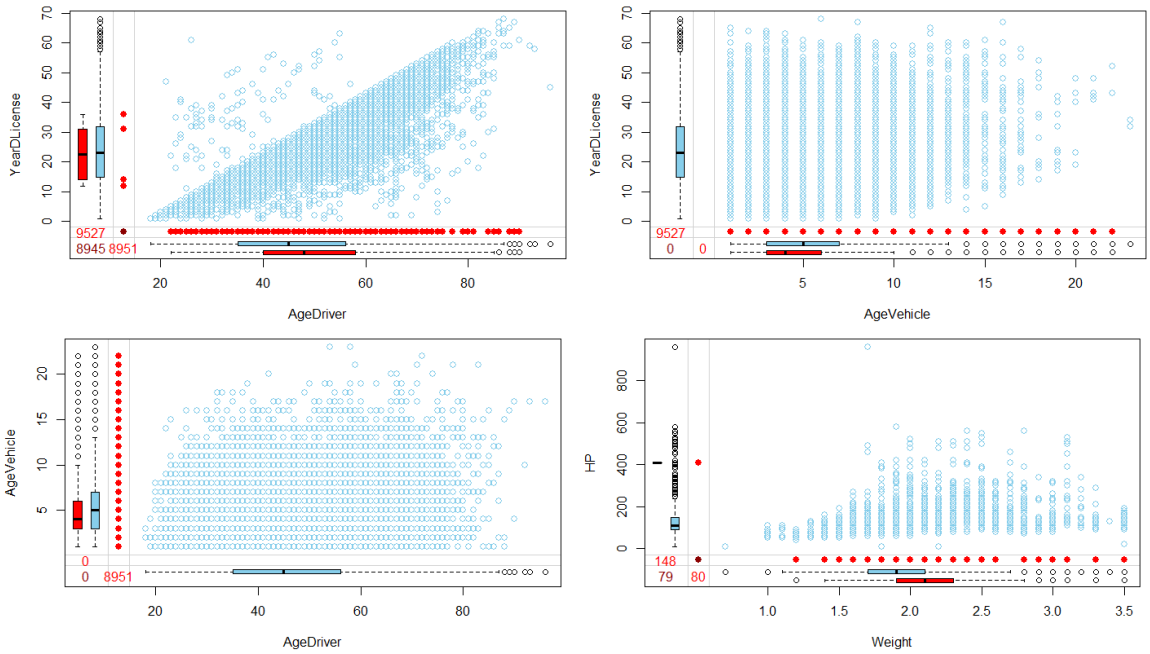


Figure 3.4: Marginplots for some combinations of database's continuous variables.

The plots in 3.4 allow us to draw some conclusions. For example, the first plot displays an evident positive linear correlation between variables `YearDLicense` and `AgeDriver`, whereas the other plots do not show strong correlations since the points are arranged in clouds without evident forms. Also, in the last plot, because of the difference between the red and blue boxplots, we can conclude that vehicles with missing value in `HP` seem to be heavier than the remaining. In this case, because variables `HP` and `Weight` are higly positively correlated, we are led to the conclusion missing values occur for policies with higher horseporwer, which may indicate that we are in the presence of MAR.

The five continuous variables in our data set are measures of time in years, weight or

power rating. This means that all variables are strictly positive, in fact even greater or equal to 18 in the particular case of the variable `AgeDriver`. For this reason, a variable transformation must be done in order to ensure that no negative value, or value below 18 for `AgeDriver`, is imputed in these variables. The transformations considered, which will be used until the end of this chapter, are shown in figure 3.5.

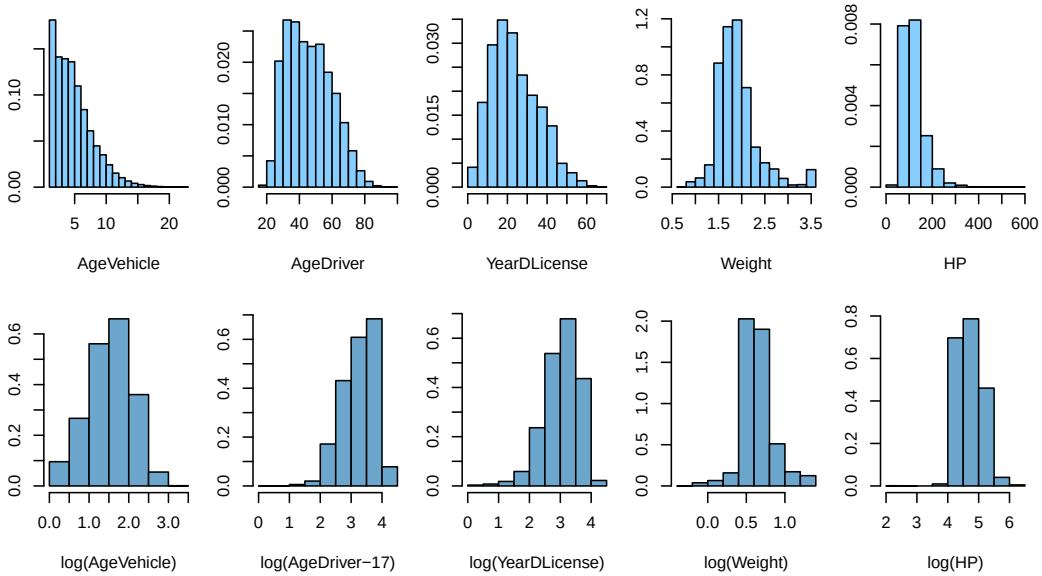


Figure 3.5: Histograms of database's continuous variables before and after transformation.

The imputing process is quite simple to execute with the **mice()** function. There are, however, two important parameters that must be taken into account:

- `m`, the number of complete data sets that will be created with imputation. The default value is set to 5;
- `method`, that can be either a single string, specifying a method to be used to all variables, or a vector of strings with length equal to the number of variables, specifying the elementary imputation method to be used for each variable. There are several methods available for implementation. The complete list of methods may be consulted at van Buuren and Groothuis-Oudshoorn (2011).

It is important to master the methods to overcome them. Because the choice of method largely influences the model's predictive ability, the method to use for each variable must be chosen wisely. Next, we present which methods were used to which variables and the main reasons for their choice:

- `norm`: Bayesian linear regression
  Bayesian linear regression is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference. Because

it assumes that the response variable is normally distributed, this method was used to impute variables `HP` and `Weight`.

- `pmm`: Predictive mean matching
  Predictive Mean Matching is a semi-parametric imputation approach for continuous variables and is especially appropriate if the normality assumption is violated. For each missing value, this method finds a set of observed values with the closest predicted mean as the missing one and imputes the missing values by a random draw from that set. Therefore, imputations are restricted to the observed values and non-linear relations can be preserved. Since variables `YearDLicense` and `AgeDriver` present skewed non-normal distributions, linear regression resulted in implausible or impossible imputed values. Therefore, `pmm` was used to impute them.

- `logreg`: Logistic regression
  Logistic regression is used to estimate the probability of binary/two-level response variables. Given that `Fuel` is the only two-level variable with missing values, it was the only variable to use this method.

- `polyreg`: Polytomous logistic regression
  This method, also known as Multinomial logistic regression, is a classification method that generalizes logistic regression to variables with more than two possible discrete outcomes. It was used to impute the three-level variable `Seats` and the four-level variable `Region`.

The final result of **mice()** was the following:

```
Multiply imputed data set
Call:
mice(data = b, m = 10, method = c("", "pmm", "pmm", "norm", "norm",
    "", "", "polyreg", "", "logreg", "polyreg", ""), seed = 2)
Number of multiple imputations:  10
Missing cells per column:
  AgeVehicle    AgeDriver YearDLicense       Weight           HP    ClassAuto
           0         8951         9527           80          148            0
       Brand        Seats          Use         Fuel       Region       CostOD
           0           25            0           13          586            0
Imputation methods:
  AgeVehicle    AgeDriver YearDLicense       Weight           HP    ClassAuto
          ""        "pmm"        "pmm"       "norm"       "norm"           ""
       Brand        Seats          Use         Fuel       Region       CostOD
          ""    "polyreg"           ""     "logreg"    "polyreg"           ""
VisitSequence:
   AgeDriver YearDLicense       Weight           HP        Seats         Fuel
           2            3            4            5            8           10
      Region
```

```
          11
PredictorMatrix:
             AgeVehicle AgeDriver YearDLicense Weight HP ClassAuto Brand Seats Use
AgeVehicle            0         0            0      0  0         0     0     0   0
AgeDriver             1         0            1      1  1         1     1     1   1
YearDLicense          1         1            0      1  1         1     1     1   1
Weight                1         1            1      0  1         1     1     1   1
HP                    1         1            1      1  0         1     1     1   1
ClassAuto             0         0            0      0  0         0     0     0   0
Brand                 0         0            0      0  0         0     0     0   0
Seats                 1         1            1      1  1         1     1     0   1
Use                   0         0            0      0  0         0     0     0   0
Fuel                  1         1            1      1  1         1     1     1   1
Region                1         1            1      1  1         1     1     1   1
CostOD                0         0            0      0  0         0     0     0   0
             Fuel Region CostOD
AgeVehicle      0      0      0
AgeDriver       1      1      1
YearDLicense    1      1      1
Weight          1      1      1
HP              1      1      1
ClassAuto       0      0      0
Brand           0      0      0
Seats           1      1      1
Use             0      0      0
Fuel            0      1      1
Region          1      0      1
CostOD          0      0      0
Random generator seed value:  2
```

So, this method creates five different data sets stored in an object of class `mids`, which has methods for some generic functions such as **summary()** or **plot()**. These data sets will be used together to pool the final result. Usually these imputed data sets are not used individually, but if one wants to have access to a certain data set, say the first, it can be done using the following code.

```
> comp1 = complete(imputed, 1)
```

In order to have access to another imputed data set, one just needs to change the second parameter.

At this point, it is possible to compare the distributions of the original and the imputed data. The following code provides two different types of plots that may be helpful for analyzing continuous variables whose missing values were imputed.

```
> densityplot(imputed)
> stripplot(imputed, pch= 20)
```

Figures 3.6 and 3.7 are the output of the first and second function, respectively. Figure

3.6 displays the density curves of the imputed data sets in magenta, and the density of the remaining values in blue. On figure 3.7, the distributions of the variables as individual points are presented to each imputation. The matching shapes and point clouds tell us that the imputed values are indeed "plausible values", i.e. that no absurd value was imputed.

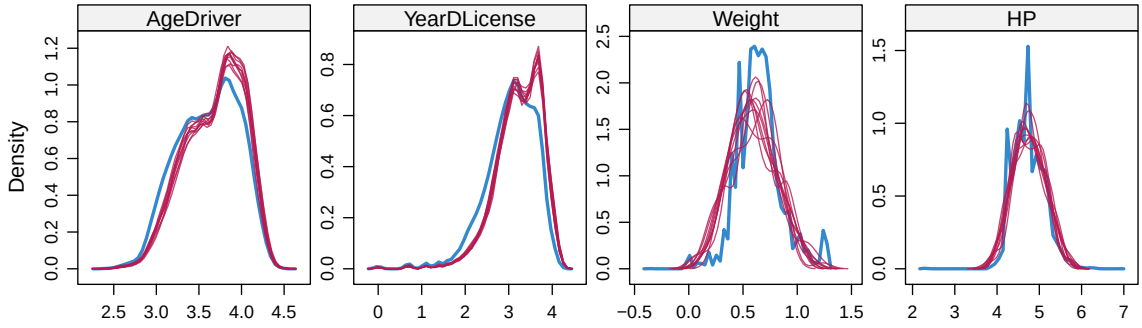


Figure 3.6: Density plots of the imputed value (magenta) and the observed values (blue) for all continuous variables with imputed values.

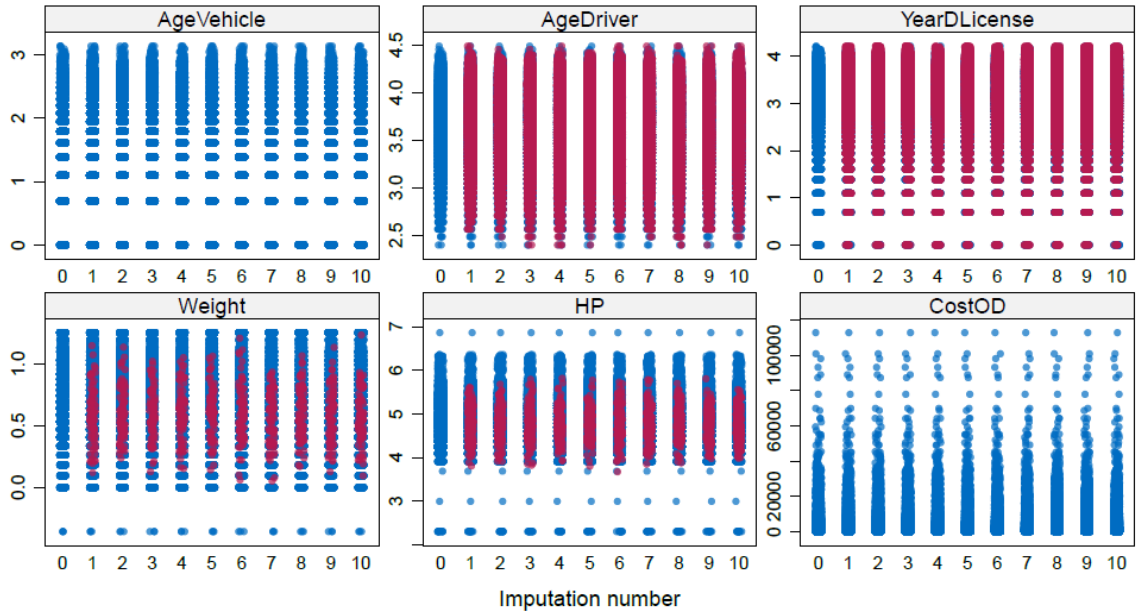


Figure 3.7: Stripes plots of the imputed value (magenta) and the observed values (blue) by imputation for all continuous variables.

The next steps, analysis and pooling, are also very clean and easy with `mice` package, and almost simultaneous. There are some analyses for which there already exists a specific function in this package, such as **glm.mids()**. However, even if there is not a specific function for the intended analysis, keeping the imputed data sets as an object

of class `mids` makes it possible to proceed by simply using the function **with()**. One way or another, the m models created can be then pooled together with the function **pool()**.

```
> models = with(imputed,
+               lm(AgeDriver ~ AgeVehicle + Weight + HP + Seats + Use + Fuel))
> modelF = pool(models)
> modelF
Call: pool(object = models)

Pooled coefficients:
(Intercept)  AgeVehicle      Weight          HP      Seats2      Seats3
 3.91960065  0.02594662  0.22868724 -0.08251346 -0.03833796 -0.06111031
       Use2       Fuel2
 0.09761945 -0.12544370

Fraction of information about the coefficients missing due to nonresponse:
(Intercept)  AgeVehicle      Weight          HP      Seats2      Seats3
  0.4281236   0.4345641   0.7263289   0.4827678   0.7472611   0.4834948
       Use2       Fuel2
  0.9018697   0.3790294
```

After pooling the results, it is possible to check the fraction of missing information obtained to each coefficient. It is important to remember that setting m too low may result in large simulation errors, especially if λ is too hight. Therefore, it is important to check if the average fraction of missing information is lined up with the number of imputations defined in table 3.1. In this case, the average λ is approximately 0.57 which means that assuming a tolerable 5% power fallout, $m = 10$ is acceptable.

3.3.1 Comparison of Variables' Gross Effects Before and After Imputation

We have seen that multiple imputation is creating complete data sets with plausible values, but it is also important to analyze if the gross effect of each variable on the cost of claims before and after imputation is similar. In fact, if a different tendency in data is detected after imputation, it might mean that the data might be MNAR and therefore multiple imputation might not be applicable.

Gross effect is the effect on the average change of the response variable of a change in the value of the predictor variable X, i.e., gross effect captures how a certain predictor variable X influences the response variable. In this case, because of our response variable, cost of claims, is a skewed distributed continuous variable (see section 5.1.2 for more details), a Gamma generalized linear model with logarithmic link function will be used to estimate the relationship between each imputed variable and the response

variable `CostOD`. For more information about generalized linear models please consult section 4.1.

Thus, the model that estimates the gross effects of a continuous variable, for example `AgeDriver`, is

$$log(\text{CostOD}) = \beta_0 + \beta_1 \times \text{AgeDriver}, \tag{3.15}$$

where β_0 is the independent regression coefficient, also called intercept, and β_1 is the coefficient of the continuous variable `AgeDriver`.

As such, the model that estimates the gross effects of variable `Seats`, for example, considering class `4-5 seats` as the reference level, is

$$log(\text{CostOD}) = \beta_0 + \beta_1 \times Z_{\text{1-3 seats}} + \beta_2 \times Z_{\text{6+ seats}}. \tag{3.16}$$

If otherwise the gross effects of a categorical variable rather than continuous are to be estimated, a set of auxiliary variables, called dummy variables, will be used. In general, a predictor variable X with $k+1$ levels may be represented by k dichotomous dummy variables Z_1, Z_2, ..., Z_k in the following way:

- Values $0, 1, 2, ...k$ are assigned to categories of X, where the reference class is assigned to 0;
- Observations that belong to level i are identified by variable Z_i, i.e, the dichotomous variable Z is defined as

$$Z_i = \begin{cases} 1 & \text{if } X = \text{level } i, \\ 0 & \text{if } X \neq \text{level } i. \end{cases} \tag{3.17}$$

Figure 3.8 presents the gross effects of untransformed continuous variables in the cost of claims. For all four plots, it is easy to verify that each variable preserves its tendency after imputation. This conclusion is especially important for variables `AgeDriver` and `YearDLicense` due to a large number of imputed cells, which was around 9000 per variable. For variables `Weight` and `HP` only 80 and 148 cells were imputed, respectively, and thus the differences between the tendency curves are so subtle that they even seem to be overlapping.

As for the categorical variables, their gross effects can be observed in table 3.2. For variables `Seats` and `Fuel`, since the number of imputed cells is only 25 and 13, respectively, coefficients after imputation vary less than 0,1% from those before imputation. As for `Region`, the higher number of imputed values result in a maximum variation of 0,84%. Notwithstanding,the obtained variations are less than 1% meaning that gross effects are similar before and after imputation.

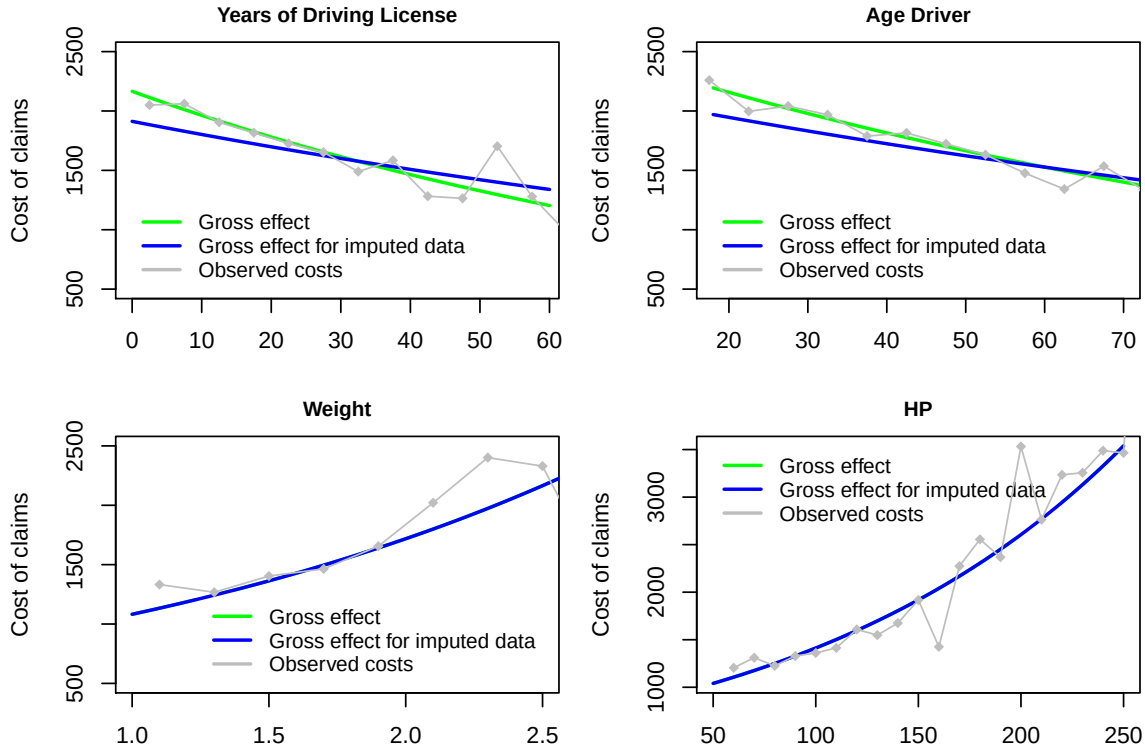


Figure 3.8: Gross effects of untransformed continuous imputed variables on cost of claims, before and after imputation.

Variable (imputed cells)	Coefficient	Before Imputation	After Imputation	Difference
Seats (25)	Intercept	1810,60	1812,15	0,09%
	1-3	0,7332	0,7333	0,01%
	4-5	1,0000	1,0000	0,00%
	6 +	0,7770	0,7762	-0,09%
Fuel (13)	Intercept	1703,48	1702,25	-0,07%
	G	1,0000	1,0000	0,00%
	D	0,9939	0,9943	0,04%
Region (586)	Intercept	1503,76	1505,50	0,12%
	North	1,2558	1,2574	0,13%
	Center	1,0000	1,0000	0,00%
	Lisbon & TV	1,0538	1,0548	0,10%
	South & Islands	0,9264	0,9341	0,84%

Table 3.2: Gross effects of categorical imputed variables on cost of claims, before and after imputation.

3.3.2 Conclusions

In principle, MICE should be able to handle large amounts of missing data. However, it is expected that variables with a high number of missing values end up with larger error terms than those with fewer, which may harm statistical inferences. Therefore, it is always necessary to analyze each variable distribution after imputation, as well as whether data trends are similar before and after imputation.

However, a preliminary analysis of which variables can be safely discarded should be done. If some variable presents a high percentage of missing values and does not seem to be statistically significant to predict the response variable, it should be excluded from the data set. Also, personal knowledge on the subject is important for selecting the included variables.

To sum up, more important than analyzing if the amount of missing data is above or below a certain "cutoff" percentage for missing data, it is to carefully know whether or not we are dealing with MNAR. Moreover, considering the intended purpose of the model, it is of major importance to carefully decide which method to use in order to impute each variable.

Chapter 4

Proposed Regression Models

A consequence of assumptions 1 and 2 made in the first chapter is that each data set observation is independent because they either concern to different policies or occur in disjoint time intervals. Therefore, regression analysis may be used to predict frequency and severity of claims. This technique is a form of predictive modeling which investigates the relationship between a dependent and independent variables.

The simplest model usually considered to predict a variable is a **linear model**, estimated by linear regression, which models the relationship between a dependent variable Y and a vector of independent variables $X = (1, X_1, ..., X_n)$ as

$$Y = \beta X^T + \epsilon \quad , \tag{4.1}$$

where $\epsilon \sim N(0, \sigma^2)$ is the random error associated to Y. The problem, however, is that linear models do not allow random deviations of Y to have distributions different from the normal. For that reason, two different types of models will be explored: generalized linear model (GLM) and generalized additive model (GAM). Both models are constructed considering three main components: (i) the distribution of the response variable, (ii) the specification of the systematic component in terms of the explanatory variables, and (iii) the link between the mean of the response variable and the systematic part. Also in this chapter, we will explore Zero-Inflated Models, a two-component mixture of GLM used for modeling count data.

4.1 Generalized Linear Models

Generalized Linear Models (GLM) were first formulated by Nelder and Wedderburn (1972) and are, as the name suggests, a generalization of linear regression. This generalization was constructed in two different directions:

- It allows the random deviations from the mean to have a distribution different from the normal. In fact, it is valid for any distribution from the exponential family, e.g. a distribution whose pdf that can be rewritten as

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{4.2}$$

  where θ is the location parameter, also called *canonical parameter*, ϕ is the dispersion parameter and $a(.)$, $b(.)$ and $c(.,.)$ known functions. If Y is a random variable whose pdf belongs to the exponential family, then $E(Y) = \mu = b'(\theta)$ and $\text{Var}(Y) = b''(\theta)a(\phi)$. Besides, sice $a(\phi)$ can be written as $a(\phi) = \frac{\phi}{\omega}$, where ω is a known constant, the equation 4.2 can also be presented as

$$f(y|\theta,\phi,\omega) = \exp\left\{\frac{\omega}{\phi}\left(y\theta - b(\theta)\right) + c(y,\phi,\omega)\right\}. \tag{4.3}$$

  The exponential family includes several well-known distributions, such as the Poisson, Binomial, Negative Binomial and Gamma distributions.

- Unlike linear regression models, in GLM, the dependent variable does not have to be estimated as a linear function of the explanatory variables; in fact, the relation between the linear predictor and the dependent variable may assume several forms.

Thereby, the mean, μ, of the dependent variable Y, which follows some exponential family distribution, depends on the independent variables X_1, X_2, .. , X_k, through:

$$\mu = E(Y) = g^{-1}(\boldsymbol{\beta}\boldsymbol{X}^T) \iff g(\mu) = \boldsymbol{\beta}\boldsymbol{X}^T \tag{4.4}$$

where $\boldsymbol{X} = (1, X_1, ..., X_k)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ is the regression parameters vector and g is a monotonic and differentiable function over the range of μ called **link function** which describes the relationship between the mean of the distribution and the linear predictor $\eta = \boldsymbol{\beta}\boldsymbol{X}^T$. There are many commonly used link functions, and their choice is informed by several considerations. However, for each distribution, there is always a well-defined canonical link function, g_c, which is the link function such that $g_c(\mu) = \theta$, where θ is the canonical parameter of the distribution. Table 4.1 on page 29 contains information about some distributions of the exponential family and its canonical link functions.

One of the reasons why GLM are so appreciated in insurance is because of their capacity of producing multiplicative models, i.e. models that can be written as

$$\mu_Y = \gamma_0 \times \gamma_1 \times \gamma_2 \times ... \times \gamma_k. \tag{4.5}$$

In this case, γ_i are the coefficients of dummy variables, also known as relativities. For example, a GLM for Poisson distribution with its canonical link function, the

Distribution	**Probability density function**	**Expected value**	**Parameter** θ	**Canonical link function**
Normal	$f(y\|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$	μ	μ	$g_c(y) = y$
Exponential	$f(y\|\lambda) = \lambda e^{-\lambda y}$	$\frac{1}{\lambda}$	$-\lambda$	$g_c(y) = -\frac{1}{y}$
Gamma	$f(y\|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$	$\frac{\alpha}{\beta}$	$-\beta$	$g_c(y) = -\frac{1}{y}$
Poisson	$f(y\|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$	λ	$log(\lambda)$	$g_c(y) = log(y)$
Binomial	$f(y\|n,p) = \binom{n}{y} p^y (1-p)^{n-y}$	np	$log\left(\frac{p}{1-p}\right)$	$g_c(y) = log\left(\frac{y}{n-y}\right)$
Negative Binomial	$f(y\|n,p) = \binom{y+n-1}{y} p^y (1-p)^n$	$\frac{np}{1-p}$	$log(p)$	$g_c(y) = log\left(\frac{y}{n+y}\right)$
Inverse Gaussian	$f(y\|\mu,\sigma) = \left(\frac{\lambda}{2\pi y^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda(y-\mu)^2}{2\sigma^2 y}\right\}$	μ	$\frac{1}{\mu^2}$	$g_c(y) = \frac{1}{y^2}$

Table 4.1: Some Exponential Family distributions and their canonical link functions.

logarithmic function, results in

$$\begin{aligned} \mu_Y &= \exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k\} \\ &= \exp\{\beta_0\} \times \exp\{\beta_1 X_1\} \times ... \times \exp\{\beta_k X_k\}. \end{aligned} \tag{4.6}$$

In actuarial statistics, a multiplicative model is usually much more plausible and easier to relate to. Imagine the following example: two clients, client A and client B, are moving from the country to the city, which is known to increase the risk of having an accident. Client A has a very expensive car and therefore pays an annual premium of 500 euros whereas client B has a small economic car and only pays 100 euros. If a multiplicative model is applied, the risk of accident increases by a fixed percentage, say 20%, meaning that client A will pay 600 euros and client B 120 euros. On the other hand, an additive model implies that the risk increases by a fixed amount, say 50 euros, which means that client A will pay 550 euros and client B 150 euros. Which model seems easier to justify?

Multiplicative models can be achieved by choosing an adequate link function, sometimes different from the canonical. For example, for a GLM with Gamma distribution, it is useful to use the logarithmic function as link function so that the final model is a multiplicative model, despite the logarithmic function not being the canonical link function. This is something that will be taken into account when formulating frequency and severity models.

4.1.1 Main Results on GLM

Because of its early development, GLM is a well known and well defined class of models. Therefore, it is easy to find good references that compile and prove the theory, such as Nelder and Wedderburn (1972), Dobson (2001) and Turkman and Silva (2000), being the last one written in Portuguese. For that reason, the present section will only state some useful definitions and results.

4.1.1.1 Parameters Estimation

The coefficients of a GLM may be estimated using the method of maximum likelihood. Consider a GLM whose pdf is defined as in 4.3, with link function g and independent variables $X_1, X_2, .., X_k$ as in 4.4. Thus, the **likelihood function** as a function of $\boldsymbol{\beta}$ defined in a set of n observations is

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} f(y_i|\theta_i, \phi, \omega_i) \\ &= \prod_{i=1}^{n} \exp\left\{\frac{\omega_i}{\phi}\left(y_i\theta_i - b(\theta_i)\right) + c(y_i, \phi, \omega_i)\right\} \\ &= \exp\left\{\sum_{i=1}^{n} \frac{\omega_i}{\phi}\left(y_i\theta_i - b(\theta_i)\right) + c(y_i, \phi, \omega_i)\right\}, \end{aligned} \tag{4.7}$$

which leads us to the **log-likelihood function** defined as

$$ln(L(\boldsymbol{\beta})) := \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(\frac{\omega_i}{\phi}\left(y_i\theta_i - b(\theta_i)\right) + c(y_i, \phi, \omega_i)\right) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}), \tag{4.8}$$

where $\ell_i(\boldsymbol{\beta})$ is the contribution of each observation y_i to the likelihood function.

The maximum likelihood estimators of $\boldsymbol{\beta}$ (MLE) are the solution of the following system of equations:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} = 0, \quad j = 0, 1, ..., k. \tag{4.9}$$

Since $\dfrac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j}$ can be expressed as

$$\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} = \frac{y_i - \mu_i}{\mathrm{Var}(X_i)} \frac{\partial \mu_i}{\partial \eta_i}\, \tilde{y}_{ij}, \quad \text{where } \frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j} = \tilde{y}_{ij}, \tag{4.10}$$

(see Turkman and Silva (2000), chapter 2) the system of equations 4.9 may be rewritten as

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathrm{Var}(X_i)} \frac{\partial \mu_i}{\partial \eta_i}\, \tilde{y}_{ij}, = 0, \quad j = 0, 1, ..., k. \tag{4.11}$$

Therefore, MLE can be obtained through the system of equations 4.11, known as likelihood equations. Usually, likelihood equations are not linear and therefore of very difficult resolution. In practice, the simplest method is to initiate a numerical optimization process for MLE and observe whether the solutions diverge or converge. The more common methods are the Newton Raphson procedure combined with Fisher Scoring and the iteratively re-weighted least squares method. As observed by Fahrmeir and Tutz (2001), the MLE may not necessarily correspond to a global maximum of the log-likelihood function. In fact, the process may stop in a local solution, so it is recommended to run a few cycles of iterations with different initial values. Nelder and Wedderburn (1972) proved that the Newton-Raphson and Fisher Scoring combined method is asymptotically equivalent to the iteratively re-weighted least squares method, since as $n \to \infty$, the distributions of the parameters become identical.

4.1.1.2 Deviance

Deviance is a goodness of fit statistic, i.e. a statistic that describes how the model fits a certain data set, generalizing the idea of using the sum of squares of residuals in ordinary least squares to all cases where model-fitting is achieved by maximum likelihood. It is constructed as a distance measure based on the likelihood ratio criterion which uses the saturated model, Ψ, to evaluate the quality of adjustment of a certain model ψ. The saturated model Ψ is the model that estimates a parameter for each observation, i.e. the model that presents greater likelihood function. The definition of **deviance** is

$$\begin{aligned} D &= -2\phi(\ell_\psi(\boldsymbol{\beta}) - \ell_\Psi(\boldsymbol{\beta})) \\ &= 2\sum_{i=1}^{n} \omega_i \Big(y_i(\theta_{\Psi i} - \theta_{\psi i}) + b(\theta_{\Psi i}) - b(\theta_{\Psi i}) \Big) \end{aligned} \tag{4.12}$$

Because D evaluates the discrepancy between the observed values and the values adjusted by the model, we have that $D \geq 0$. Hence, the greater the discrepancy, the greater will be the value of D. On the other hand, a model with a perfect fit will have $D = 0$, as is the case of the saturated model. Note that deviance is defined to be independent of ϕ.

4.1.1.3 Residuals

As pointed out by Nelder and Wedderburn (1972), residuals can be used to explore the adequacy of fit of a model in respect of choice of variance function, link function and terms in the linear predictor, and they may indicate the presence of anomalous values requiring further investigation. Residuals are, in a simple approach, the difference

between the observed value and the estimated value of the quantity of interest. They are very easily understood in the Normal case since it is possible to write

$$y_i = \hat{\mu}_i + (y_i - \hat{\mu}_i), \tag{4.13}$$

or equivalently, observed value = fitted value + residual. For GLM however, the formula is not so linear and a generalization is necessary in order to make it applicable for all distributions of the exponential family.

Several definitions of residuals have been proposed and **Pearson's residuals** is one of the most used. For each observation, it is defined as

$$R^P{}_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\text{Var}(\hat{\mu}_i)}}, \tag{4.14}$$

which should have approximately zero mean and variance ϕ, if the model is correct. These residuals should not display any trend in mean or variance when plotted against the fitted values, or any covariates (whether included in the model or not).

Despite being one of the most used methods, Pearson's residuals only work for normally distributed data, leading to very asymmetrical distributions otherwise. For that reason, the **deviance residuals** are often preferable. Once again, because deviance is for GLM as residual sum of squares for ordinary linear models, it is possible to define the residuals as the square root of the components of the deviance with the appropriate sign attached. So, writing the equation 4.12 as $D = \sum_{i=1}^{n} d_i$, the deviance residuals can be defined as

$$R^D{}_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}, \quad \text{where } \sum_{i=1}^{n} \left(R^D{}_i\right)^2 = D. \tag{4.15}$$

4.1.1.4 R^2, Pearson's Pseudo R^2 and Relative Absolute Error

The **coefficient of determination,** R^2 , is an output of linear regression analysis, used to measure the proportion of the variance in the dependent variable that is predictable from the independent variable. This coefficient is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{4.16}$$

where y_i are the observed values, $\hat{\mu}_i$ the predicted values by the linear regression, $\bar{y}$ the mean y_i value and N the sample size. Because the coefficient of determination is the square of the correlation between predicted values and the observed values, it ranges from 0 to 1, 0 meaning that the dependent variable cannot be predicted from

the independent variable and 1 meaning that the dependent variable can be predicted without error from the independent variable.

The problem with this coefficient however is that it assumes that residuals follow a normal distribution and therefore it can only be applied in case of a linear regression. As a consequence, several authors have developed other measures, known as pseudo R^2, based on the coefficient of determination in order to generalize this idea and in particular to make them applicable to GLM. One of those measures is based on Pearson's residuals and therefore called **Pearson's Pseudo** R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{\sqrt{\text{Var}(\mu_i)}}}{\sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{\sqrt{\text{Var}(y)}}} \tag{4.17}$$

Another useful measure of goodness-of-fit is the **Relative Absolute Error**, which takes the total absolute error and normalizes it by dividing by the total absolute error of the mean:

$$RAE = \frac{\sum_{i=1}^{N} |y_i - \hat{\mu}_i|}{\sum_{i=1}^{N} |y_i - \bar{y}|}, \tag{4.18}$$

For a perfect fit, the numerator is equal to 0 and therefore $RAE = 0$. If, however, the mean predicts the observed values better than the model, the numerator will be greater than the denominator and $RAE > 1$. Therefore, the range of this measure is from 0 to infinity.

4.1.1.5 Hypothesis Testing

Some of the most used hypothesis tests in statistics are those which focus on one or more regression parameters. These tests evaluate whether model explanatory variables are significant, i.e. they measure the relevance of including a coefficient β in the model. The models presented next are equivalent and both have univariate and multivariate versions. The hypothesis is formulated as follows:

- $H_0 : \beta = 0$,
- $H_1 : \beta \neq 0$.

Notice that testing H_0 is equivalent to comparing the goodness-of-fit of models with $\beta = 0$ (M_0) and with $\beta \neq 0$ (M_1).

The first test is the **log-likelihood ratio statistic (LR)**, which expresses how many times more likely the data are under one model than the other. Let $\hat{\beta}_0$ be the maximum

likelihood estimator of β in M_0 and $\hat{\beta}_1$ the maximum likelihood estimator of M_1. Then:

$$\text{LR}(M_0, M_1) = -2\left(\ell(\hat{\beta}_0) - \ell(\hat{\beta}_1)\right) = \frac{D_{M_0} - D_{M_1}}{\phi}. \tag{4.19}$$

High values of LR indicate that models M_0 and M_1 are very different, which means that H_0 is inappropriate. Also, if the number of observations is high, it can be used to compute a p-value in order to formally decide whether or not to reject H_0, since LR asymptotically follows a χ^2 distribution:

$$\text{LR}(M_0, M_1) = \frac{D_{M_0} - D_{M_1}}{\phi} \sim \chi^2(1). \tag{4.20}$$

Wald's test is a quadratic approximation of log-likelihood ratio statistic. The univariated version of the test states that, under H_0 we get:

$$W = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim N(0,1), \quad j = 0, 1, ...p. \tag{4.21}$$

The multivariate version of this test is computationally heavy and therefore not very much used, but the univariate version here presented is. In fact, for some unjustified reason, this method tends to be more used than the log-likelihood ratio statistic. The two test are equivalent and the results usually are very similar, especially if the number of observations is high enough.

The **Vuong test** (Vuong (1989)), is a statistical test that compares the predicted probabilities of two models. The models can be nested or non-nested. Basically, it states that under the null hypothesis H_0, the two model M_1 and M_2 fit equally well, i.e. the expected value of their log-likelihood ratio equals zero, which means that the asymptotic distribution of the log-likelihood ratio statistic, LR, is normal:

- $H_0 : \text{LR}(M_1, M_2) = \ell_{M_1} - \ell_{M_2} = 0 \Longrightarrow \dfrac{\text{LR}(M_1, M_2)}{\text{V(LR)} \times \sqrt{n}} \longrightarrow N(0,1)$
- $H_1 : \text{LR}(M_1, M_2) \neq 0$

The tests here presented will be used during the analysis as tools improve and decide which models are preferable.

4.1.1.6 AIC and BIC

The **Akaike information criterion (AIC)**, developed by Akaike (1974) and the **Bayesian information criterion (BIC)**, developed by Schwarz (1978), are measures based on log-likelihood which are used to compare the information provided by two models, even in case they are non-nested. The log-likelihood is itself a measure of

information of the model, but as it increases with the number of estimated parameters a correction must be introduced in order to avoid overfitting the data. Therefore, AIC and BIC are defined as:

$$AIC_M = -2\ell_M + 2p_M, \tag{4.22}$$

$$BIC_M = -2\ell_M + 2p_M \times log(n), \tag{4.23}$$

where ℓ_M is the model log-likelihood, p_M the number of parameters estimated in the model and n the number of observations. As such, in both cases, the lower the value the better the model.

However, it is important to notice that both AIC and BIC do not provide hypothesis tests for model comparison and their values do not carry any information on the quality of the model by themselves. Therefore, if all candidate models fit poorly, these measures will not tell anything about that.

4.1.2 Overdispersion in Poisson GLM

Overdispersion means that the variability of the data is larger than the mean and is, according to Hilbe (2014), *the foremost problem facing analysts who use Poisson regression when modeling count data.* Hilbe (2014) states that there are two main reasons for overdispersion, which give rise to apparent overdispersion and real overdispersion. Apparent overdispersion is due to missing covariates or interactions, outliers in the response variable, non-linear effects of covariates entered as linear terms in the systematic part of the model or choice of the wrong link function. A very good example of apparent overdispersion is presented by Hilbe (2014) when he simulates a Poisson variable using five explanatory variables and then applies a Poisson model using only two explanatory variables, creating overdispersion. Real overdispersion exists when any of the previously mentioned causes occur. This may append because of the variation in the data really is larger than the mean or even because clustering of observations, correlation between observations or the existence of too many zeros, even though it may or may not cause overdispersion.

The most common way of detecting overdispersion is based on the χ^2 approximation of residual deviance. Basically, if there is overdispersion, D/ϕ follows some χ^2 distribution with $n - p$ degrees of freedom and therefore an estimator of ϕ can be defined as

$$\hat{\phi} = \frac{D}{n-p}. \tag{4.24}$$

Therefore, if $\hat{\phi}$ is about 1, it is safe to assume that there is no overdispersion. If however the ratio is larger enough than 1, there is evidence of overdispersion which will not be taken into account by the model, invalidating the results. The problem with Poisson

GLM with overdispersion is that the mean and the variance of the response variable are no longer equal, thus a relationship between the two measures must be explicitly specified.

So, how much larger than 1 should $\hat{\phi}$ be before we need to make a correction? The answer is not direct, because it depends on the significance of the parameters. First, it is important to know that by introducing a dispersion parameter in the model, the standard errors of the parameters are multiplied by the square root of ϕ. So, basically, if the parameters of a Poisson GLM are highly significant, a ratio greater than 1 but not that high, e.g. 1.5, will not be a problem. If however there is a parameter that is almost not significant, e.g. with a p-value of 0.03, then multiplying the standard error by the square root of 1.5 may change the p-value in something that is no longer significant.

In general a $\hat{\phi} > 1.5$ means that some action needs to be taken to correct it and the first attempt should be using a quasi-Poisson distribution. If however $\hat{\phi}$ is larger than 15 or 20, the negative binomial distribution must be considered (Zuur et al. (2009)).

Hoef et al. (2007) explain the difference between quasi-Poisson and negative binomial distributions as follows: for the response variable Y, if $E(Y) = \mu$, then quasi-Poisson assumes $\text{Var} = \phi\mu$, where ϕ is the dispersion parameter and $Y \sim \text{quasi-P}(\mu, \phi)$, while negative binomial assumes $\text{Var}(Y) = \mu + \alpha\mu^2 = \mu(1 + \alpha\mu)$, where α is the shape parameter of negative binomial distribution and $Y \sim \text{NB}(\mu, \alpha)$. This means that, in the first case, model formulation has the advantage of leaving parameters in a natural, interpretable state and allows standard model diagnostics without a loss of efficient fitting algorithms. In the second case, the overdispersion (the amount in excess of μ) is the multiplicative factor $1 + \alpha\mu$, which depends on μ. Even more, it is important to notice that for quasi-P(μ, ϕ) the variance is linearly related to the mean, whereas for NB(μ, α) the variance is quadratic in the mean.

4.2 Zero-Inflated Models

Zero-inflated models (ZI) are two-component mixture models that combine a point mass at zero with a count distribution. Basically, the two components correspond to two zero generating processes, being one governed by a binary distribution that generates structural zeros and the other governed by a count distribution, generating counts that may be zero. This means that there is an overlapping estimation of zeros, which make it impossible to estimate the models separately. Thus, zero-inflated model is estimated as

$$P(Y = k) = P(\text{Bin} = 0) \times I_{k=0} + P(\text{Bin} > 0) \times P(\text{Count} = k), \tag{4.25}$$

where

$$I_{k=0} = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4.26}$$

According to Zuur et al. (2009), zeros generated by the count model are defined as *true zeros* and zeros generated by the binary process are *false zeros*. In the context of claims count, the former can be interpreted as the true absence of claims while the last as the non-reported claims. Figure 4.1 illustrates this idea.

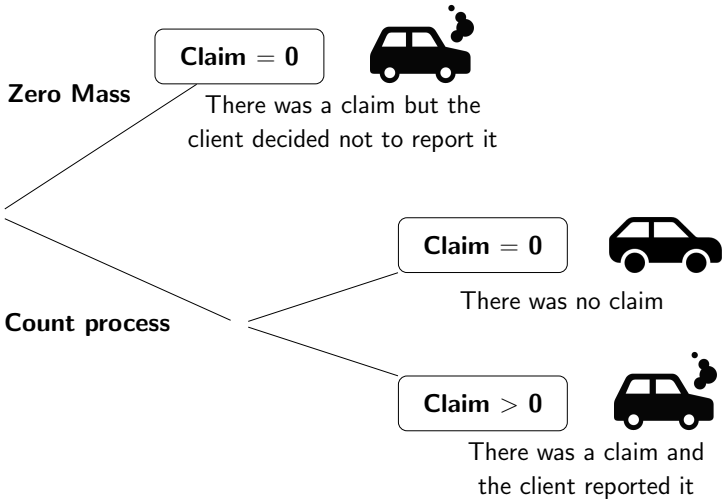


Figure 4.1: Sketch of the underlying principle of ZI models (Figure inspired in Zuur et al. (2009)).

Usually, for the binary prediction, a binomial distribution is usually used with logit link functions, i.e. a logistic regression, which is assumed to occur with probability π_i. The relation between pi_i and the linear predictor is then given by

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{\gamma}\boldsymbol{Z}_i \ , \tag{4.27}$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, ..., \gamma_k)$ is the vector of coefficients of dimension $k+1$ associated to the k variables $\boldsymbol{Z} = (Z_1, Z_2, ..., Z_k)$.

As for the count model, which occurs with probability $1 - \pi_i$, usually a Poisson or negative binomial regression with logarithmic link function is used, as well as the geometric distribution which is a particular case of the negative binomial with size parameter set to 1. In this case, we have that

$$\log(\mu_i) = \boldsymbol{\beta}\boldsymbol{X}_i \ , \tag{4.28}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_q)$ is the vector of coefficients of dimension $q+1$ associated to the q variables $\boldsymbol{X} = (X_1, X_2, ..., X_q)$.

If the Poisson distribution is used, the model is called **Zero-Inflated Poisson (ZIP)**. Otherwise, if negative binomial distribution is used, the model is called **Zero-Inflated Negative Binomial (ZINB)**.

4.2.1 Zero-Inflated Poisson Models

So, as explained before, zero-inflated Poisson models the mean of a Poisson variable, μ_i, through a Poisson regression and models false zeros which occur with probability π_i through logistic regression. Thus, this model can be expressed in terms of μ_i and π_i as:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\mu_i}\,, & \text{if } y_i = 0 \\ (1-\pi_i)\dfrac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\,, & \text{if } y_i > 0 \end{cases} \tag{4.29}$$

Cameron and Trivedi (2013) prove that the mean and variance of a ZIP model are:

$$E(Y_i) = \mu_i(1-\pi_i), \tag{4.30}$$

$$\text{Var}(Y_i) = \mu_i(1-\pi_i)(1+\pi_i\mu_i). \tag{4.31}$$

Note that if the probability of false zeros is zero, that is $\pi_i = 0$, we obtain the mean and variance equations from the Poisson GLM. Otherwise, if $\pi_i > 0$, then the variance is larger than the mean, which means that the excessive number of zeros may be a cause for overdispersion.

4.2.2 Zero-Inflated Negative Binomial Models

Zero-inflated Negative Binomial models are very similar to zero-inflated Poisson models and must be used when data presents not only too many zeros but also high overdispersion.

Negative Binomial pdf is usually described as a function of n and p, the number of trials and the probability of success in each trial, respectively, as presented in table 4.1. Notice that

$$\binom{y+n-1}{y} = \frac{\Gamma(y+n)}{y!\,\Gamma(n)}, \quad \text{where } \Gamma(y+n) = (y+n)!. \tag{4.32}$$

However, in negative binomial regression, the distribution is specified in terms of its mean μ (Hilbe (2011)). The result is the following probability density function:

$$f(y_i|\mu_i, k) = \frac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)} \times \left(\frac{k}{k+\mu_i}\right)^k \times \left(1-\frac{k}{k+\mu_i}\right)^{y_i}. \tag{4.33}$$

Thus, ZINB model can be expressed in terms of μ_i and π_i as:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\dfrac{k}{k+\mu_i}\right)^k, & \text{if } y_i = 0 \\ (1-\pi_i) \times \dfrac{\Gamma(y_i+k)}{\Gamma(k)\Gamma(y_i+1)}\left(\dfrac{k}{k+\mu_i}\right)^k \left(1-\dfrac{k}{k+\mu_i}\right)^{y_i}, & \text{if } y_i > 0 \end{cases} \tag{4.34}$$

where k is an overdispersion parameter assumed not to depend on covariates. Therefore, the mean and variance of ZINB are (Zuur et al. (2009)):

$$E(Y_i) = \mu_i(1-\pi_i), \tag{4.35}$$

$$\text{Var}(Y_i) = \mu_i(1-\pi_i)\left(\mu_i + \frac{{\mu_i}^2}{k}\right) + {\mu_i}^2({\pi_i}^2 + \pi_i) \tag{4.36}$$

Notice that, if $\pi_i = 0$ we obtain the variance of a Negative Binomial distribution and, moreover, if k is large enough when compared with μ_i^2, then the term μ_i^2/k approximates zero and variance equals μ_i, which means that the distribution converges to a Poisson distribution. So, the smaller k, the larger the overdispersion.

4.3 Generalized Additive Models

Generalized Additive Models (GAM) were developed by Hastie and Tibshirani (1990) with the aim of combining the properties of generalized linear models and additive models. GAM extend GLM by replacing the linear form $\beta_0 + \sum_j \beta_j X_j$ with the additive form $\beta_0 + \sum_j f_j(X_j)$, creating a semi-parametric model that can be written as

$$\begin{aligned} \mu = E(Y) &= g^{-1}(\beta X^T + f_1(X_1) + f_2(X_2) + f_3(X_3, X_4) + ...) \\ \iff g(\mu) &= \beta X^T + f_1(X_1) + f_2(X_2) + f_3(X_3, X_4) + ... \end{aligned} \tag{4.37}$$

where the dependent variable Y follows some exponential family distribution, βX^T is the strictly parametric part of the model such as the linear predictor in GLM, and f_j are smooth functions of the independent variables $X_1, X_2, .., X_k$, the non-parametric part of the model.

GAM are a valuable tool in modeling because of the flexibility they introduce in the relation between the dependent variable Y and the covariates. Using smooth functions rather than detailed parametric relationships gives us the possibility of avoiding inexplicable oscillations in the model and capturing data trends. However, this leads us to two new questions: how to represent the smooth functions and how smooth they should be.

4.3.1 An Introduction to Smooth Functions

Let us start by considering the simple case of a model with one smooth function of one covariate,

$$Y = f(X) + \epsilon \tag{4.38}$$

where Y is the response variable, f a smooth function of the covariate X and $\epsilon \sim N(0, \sigma^2)$ is the random variation. To further simplify matters, suppose that X lies in the interval $[0, 1]$.

In order to use the same techniques as in GLM, it is essential that 4.38 becomes a linear model which may be done using a *basis* for $f(X)$. This means that if $b_i(X)$ is the i-th element of a basis of length m for the space of f, then f can be written as

$$f(X) = \sum_{i=1}^{m} b_i(X)\beta_i. \tag{4.39}$$

Combining equations 4.38 and 4.39, we get a linear model as desired:

$$Y = \sum_{i=1}^{m} b_i(X)\beta_i + \epsilon. \tag{4.40}$$

The problem here is that we do not know in reality the values of β_i, and not only a small fluctuation of these values might result in very distinct smooth curves but also its values are not flexible enough to model more complicated patterns. The idea is then to divide the X values into n segments and fit the model 4.40 using ordinary least squares on each segment. The result is a more complicated pattern with known betas per segment. This may, however, create discontinuities at $n - 1$ points where the lines come together, known as knots. Even more, we face the problem of choosing the number of knots and their location.

4.3.1.1 Splines

Almost all the smooths considered in the literature are based in some way on splines and therefore it is worth spending some time on its theoretical properties. Splines are numerical piecewise-defined function known for their interpolation capacities, often preferred to polynomial interpolation for being able to approximate more complex shapes.

The most commonly used splines are **cubic splines**. Given a set of points (knots) $\{(x_i, y_i) \in \mathbb{R}^2 : i = 1, ..., n \wedge x_i < x_{i+1}\}$, the cubic spline interpolating these points is a function made up of sections of cubic polynomials, one for each $[x_i, x_{i+1}]$, continuous

to second derivative and that has zero second derivatives at the end knots x_1 and x_n. Even more, its natural form, $s(x)$, is the one function that minimizes

$$J(f) = \int_{x_1}^{x_n} f''(x)^2 dx. \tag{4.41}$$

Green and Silverman (1994) proved that any different function from $s(x)$ in the above-mentioned conditions will have a higher value for J, which defines a sense in which the natural cubic spline is the smoothest possible interpolation through any set of data. Boor and Golub (1978) also presented a number of results demonstrating that cubic splines are, in several approaches, the best achievable approximation. These results suggest that splines may provide a satisfactory representation for smooth curves in statistical models. Whatever the true underlying a smooth function is, a spline may approximate it closely.

So that a cubic spline can be defined, first we must define a basis. Given the location of the n knots, there are many ways of writing down a basis for cubic splines. A simple approach for a basis of dimension $n+2$ is presented in Gu (2013), defined as:

$$b_1(x) = 1, \quad b_2(x) = x \quad \text{and } b_{i+2} = R(x, x_i^*) \quad i = \{1, ..., n\} \tag{4.42}$$

where x_i^* is the location of the i-th knot and $R(x, z)$ is

$$\begin{aligned} R(x,z) = &\left(\left(z - \frac{1}{2}\right)^2 - \frac{1}{12}\right) \times \left(\left(x - \frac{1}{2}\right)^2 - \frac{1}{12}\right) \times \frac{1}{4} - \\ &- \left(\left(|x-z| - \frac{1}{2}\right)^4 - \frac{1}{2}\left(|x-z| - \frac{1}{2}\right)^2 + \frac{7}{240}\right) \times \frac{1}{24} \end{aligned} \tag{4.43}$$

So, having defined the basis, the regression parameters β_i can be estimated by ordinary least squares.

Recall that the b_i depend on the number of knots, and the number and position of knots have a powerful effect on the smoothness of the final spline. Generally, the more knots we use, the less smooth the curve becomes. This means that if in one hand sufficient knots are needed in order to capture the tendency of the observations, on the other hand too many knots compromise the smoothness degree. Therefore, the next question is how many knots to use in order to control the degree of smoothness. Keele (2008) gives a general recommendation to use 3 knots if there are less than 30 observations and 5 knots if there are more than 100 observations. However, in insurance data sets we usually face thousands or even millions of observations and therefore these recommendations might not be suitable. As to the placement of the knots, this is typically done using quartiles or equidistant positions.

4.3.1.2 Controlling the Degree of Smoothing with Penalized Regression Splines

The number of knots considered can affect the degree of smoothing. Therefore, the first approach would be trying to achieve the optimal number of knots by hypothesis testing procedures or backward selection methods. However, because a model based on $k-1$ evenly spaced knots will not generally be nested within a model based on k evenly spaced knots, such approach can be extremely complicated. Other methods, such as starting with a fine grid of knots and simply drop knots sequentially, may result in uneven knot spacing, leading to poor model performance.

Wood (2006) proposes an alternative to controlling smoothness based on regression splines with a penalized component. Instead of focusing on the optimal number of knots, the idea is to keep the basis dimension fixed, large enough, and controlling the smoothness of the model by adding a penalization for *wiggliness* to ordinary least squares.

Linear regression implies that we can find the parameters β in equation 4.40 by minimizing the sum of squares SS, which may be written in matrix notation as

$$SS = \sum_{j=1}^{n}(Y_i - \mu_i)^2 = ||Y - \mu||^2 = ||Y - \boldsymbol{\beta X}||^2, \tag{4.44}$$

where $||.||$ stands for the Euclidean norm, Y contains all the observed data in vector format, β all the parameters in vector format, and X all the b_i of the equation 4.40. Wood argues that rather than fitting the model by minimizing S, the model could be fit by minimizing

$$||Y - \boldsymbol{\beta X}||^2 + \lambda \int_0^1 f''(x)^2 dx, \tag{4.45}$$

where λ is the smoothing parameter. The second-order derivative indicates how smooth the curve is, which means that a high value of $f''(x)$ indicates that f is highly non-linear and $f''(x) = 0$ produces a straight line. Hence, for a high value of λ, the penalty for having a non-smooth curve is large resulting in a straight line, whereas a small value of λ produces a low penalty, probably creating a less smooth curve.

Because $\int_0^1 f''(x)^2 dx = \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}$ (see exercise 7, chapter 3, Wood (2006)), where $\boldsymbol{S}$ is the coefficiens matrix that can be expressed in terms of $b_i(X)$, it is possible to define the penalization function with matrix notation as

$$\mathcal{P}(\boldsymbol{\beta}; \lambda) = ||Y - \boldsymbol{\beta X}||^2 + \lambda \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta}. \tag{4.46}$$

Since $\mathcal{P}(\boldsymbol{\beta}; \lambda)$ is a positive not-limited function, the value $\hat{\boldsymbol{\beta}}$ at which $\mathcal{P}$ attains a

minimum is the zero of the first derivative:

$$\begin{aligned}\frac{\partial\, \mathcal{P}(\boldsymbol{\beta};\lambda)}{\partial \boldsymbol{\beta}} &= 2\,\frac{\partial\,(Y-\boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\,(Y-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\,\frac{\partial\,(\boldsymbol{\beta}^T\boldsymbol{S}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= -2(\boldsymbol{X}^T Y - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{S}\boldsymbol{\beta}.\end{aligned} \tag{4.47}$$

So, equaling the previous equation to zero, we get the value which minimizes equation 4.46, the penalized least squares estimator of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{S})^{-1}\boldsymbol{X}^T Y. \tag{4.48}$$

Therefore, Woods' alternative drops the idea of looking for the optimal number of knots and focus on minimizing the equation 4.46, also called penalized least squares. The advantage of this method is that, provided that the basis dimension is larger than expected, neither the choice of the basis nor the exact location of the knots will be of great influence in the model fit. The problem, however, is that this method implies that a value for λ has to be chosen, and the choice of λ determines the model flexibility and ultimately the shape of the curve. Thereby, the question to be asked now is what is the optimal value for λ.

Choosing the Smoothing Parameter

The parameter λ severely influences the degree of smoothness, creating over smoothness if λ is too high and under smoothness if λ is too low. This means that choosing a value for λ before minimizing the equation 4.46 may originate a spline estimate $\hat{f}$ that may not be close to the true function f.

The criterion developed by Wood to ensure that $\hat{f}$ is as close as possible to the real function f is to choose a value for λ that minimizes

$$M = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(x_i) - f(x_i)\right)^2. \tag{4.49}$$

The problem is that being f an unknown function, it is not possible to use M directly. However, it is possible to estimate $E(M) + \sigma^2$, the expected squared error of M, by *cross validation*.

The *ordinary cross validation score* is defined as follows. Firstly, the observation i is dropped, the smoother is estimated using the remaining $n-1$ observations and the value for observation i is predicted from the estimated smoother. Then, the squared difference between the predicted and real value is calculated. These squared differences are then averaged for all n observations:

$$\nu_0 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(x_i)^{[-i]} - Y_i\right)^2, \quad Y_i = f(x_i) + \epsilon_i. \tag{4.50}$$

Since ϵ_i and $\hat{f}(x_i)^{[-i]}$ are independent and $\epsilon_i \sim N(0, \sigma^2) \Rightarrow \epsilon_i^2 \sim \chi(\sigma^2)$, the expected value of ν_0 can be simplified as

$$\begin{aligned} E(\nu_0) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(x_i)^{[-i]} - Y_i\right)^2\right) \\ &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(x_i)^{[-i]} - f(x_i) - \epsilon_i\right)^2\right) \\ &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left[\left(\hat{f}(x_i)^{[-i]} - f(x_i)\right)^2 - 2\left(\hat{f}(x_i)^{[-i]} - f(x_i)\right)\epsilon_i + \epsilon_i^2\right]\right) \\ &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(x_i)^{[-i]} - f(x_i)\right)\right) + \frac{1}{n}E\left(\sum_{i=1}^{n}\epsilon_i^2\right) \\ &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(x_i)^{[-i]} - f(x_i)\right)\right) + \sigma^2. \end{aligned} \tag{4.51}$$

If the sample is large enough, which is usually the case in insurance data sets, it is reasonable to assume $\hat{f}(x_i)^{[-i]} \approx \hat{f}(x_i)$, which means that

$$E(\nu_0) \approx E(M) + \sigma^2. \tag{4.52}$$

Thus, since it is not possible to choose λ minimizing M, it is reasonable to choose λ minimizing ν_0. This method is known as ordinary cross validation.

4.3.1.3 Two More Things About Splines

More than one spline at a time

It is common to create GAM with more than one smooth function. In fact, not only it is usual to have more than one smooth function in the model but also it is possible to have a smooth function for a set of variables. The last case is particularly useful analyzing geographical data where it is of major interest to create a smooth function for latitude and longitude together. We will not explore this case.

So, consider a model with two smooth functions

$$Y = f_1(X_1) + f_2(X_2) \tag{4.53}$$

where

$$f_1(X_1) = \sum_{i=1}^{m}\beta_i\, b_i(X_1) \quad \text{and} \quad f_2(X_2) = \sum_{i=1}^{m}\beta_i\, b_i(X_2). \tag{4.54}$$

Using two smooth functions instead of one has an effect on the definitions of the Y, $\boldsymbol{X}$ and $\boldsymbol{\beta}$ in equation 4.44, but the essence is basically the same. In this case, the

penalization function becomes

$$\mathcal{P}(\boldsymbol{\beta}; \lambda_1, \lambda_2) = ||Y - \boldsymbol{\beta}\boldsymbol{X}||^2 + \lambda_1 \int_0^1 f_1''(x)^2 dx + \lambda_2 \int_0^1 f_2''(x)^2 dx. \tag{4.55}$$

This generalization of the penalization function allows different amounts of wiggliness per smoothing spline, which means that some smoothers can be smooth whereas others are not.

Other types of splines

The family of splines is rather large. In section 4.3.1.1 **cubic splines** were presented and even though they seem to be somewhat ideal smoothers, they have as many free parameters as there are data to be smoothed. Actually, the large number of parameters for univariate smoothing with cubic splines is not that problematic, but as soon as we try to deal with more covariates the computational expense becomes severe. An obvious compromise between retaining the good properties of splines and computational efficiency is to use penalized regression splines as introduced in section 4.3.1.2.

Cubic regression splines differ from cubic splines on how the basis is defined. Instead of being defined as in equation 4.42, this approach parameterizes the spline in term of its values at the knots. The advantage is that this basis does not require any rescaling of the predictor variables. **Cyclic cubic regression splines** are constructed in the same way but carry the extra property of being cyclic, i.e. the function has the same value and first few derivatives at its upper and lower boundaries. This type of splines is appropriate for several cycle variables, being mostly used for smoothing time effects.

The splines covered so far are useful in practice. Notwithstanding, not only they imply the choice of knots locations, which introduces an extra degree of subjectivity into the model fits, but also they are only useful for representing smooths of one predictor variable. **Thin plate splines** were formulated by Duchon (1977) and are a solution to these problems, making thin plate splines probably the best solution to deal with more than one predictor variable. The only problem seems to be their computational cost, because these smoothers have as many unknown parameters as there are data and this method implies the decomposition of the matrix of parameters into its eigenvalues and eigenvectors. Fortunately, the method of Lanczos iteration (see Lanczos (1950)) may be employed to find them using substantially lower cost operations.

4.3.2 Final Notes About GAM

As seen before, creating a generalized additive model goes through choosing a basis for the smooth functions, estimating smoothing parameters and model coefficients for a penalized likelihood maximization problem together with associated measures of function wiggliness. In section 4.3.1, all results were presented using the Normal distribution because it simplifies the matter and keeps smooth functions easy to calculate. However, the theory behind GAM supports a generalization for any distribution of the exponential family, which makes it far more complex than what was presented. Because the extension of the methods used before result in high computational costs, several iterative and numerical methods are used in order to estimate all necessary parameters. For those who wish to know more about this matter, all the theory behind GAM is highly detailed in chapter 4 of Wood (2006).

Chapter 5

Risk Models Using R

As explained in chapter 1, the idea behind risk models is to classify and differentiate the risks and, ultimately, calculate the pure premium for each one of those risks. To do it so, it is necessary to estimate the frequency and severity of claims for each risk, since

$$\text{pure premium} = \text{claims frequency} \times \text{claims severity}. \tag{5.1}$$

The different nature of the models leads us to study them separately. In fact, frequency models are a *count data* problem whereas severity deals with costs of claims, a continuous variable. This means that each model has to be handled separately. However, it is important to notice that this approach is only valid under the assumption that frequency and severity are independent.

This chapter is developed as follows. Section 5.1 starts with the construction of standard frequency and severity models by insurance companies, with a Poisson GLM for frequency and a Gamma GLM for severity. With those models, a complete tariff is created and interpreted. The aim of this tariff is to be used as a reference to further models. Next, in sections 5.2 and 5.3, some problems such as overdispersion and data having too many zeros will be analyzed and tested. Alternative distributions and link functions will be used in order to find out which option best suits our case. In section 5.4, smooth functions are introduced to capture time effects and the tendency of continuous covariates in both frequency and severity models. Finally, in section 5.5, a tariff is constructed using the previously explored methods which proved to be useful. In order to evaluate the utility of these methods, the classical tariff and the final tariff are tested against each other.

5.1 Classical Approach

Although insurance companies claim to be one of the private sectors that most valuable statistics, much of the actuarial work involves little or no statistic. Moreover, when in fact some statistical methods are used, they are hardly innovative or about hot topics. The truth is that insurance companies, when it comes to statistics, are very cautious, preferring sometimes to rely on the (often fallacious) experience of insurance portfolio managers than to rely on rigorous mathematical techniques.

Creating a tariff is one of the few topics where an actuary really has the opportunity to deal with statistics. Still, this interaction is usually controlled and very limited. In fact, most insurance companies around the world buy software designed specifically for this purpose, such as the EMBLEM software, developed by *Towers Watson*[1]. This type of software is user-friendly but highly limiting, usually only allowing to create two types of models - a Poisson GLM for claims frequency and a Gamma GLM for severity.

The aim of this section is to reproduce what is usually done nowadays in insurance companies around the world to estimate pure premium. We will study the particular case of a motor insurance policies, creating separate frequency and severity models.

5.1.1 Frequency Models

A model for frequency of claims consists in a model capable of predicting the number of claims that a client, here represented as a policy, may have during a time interval. Therefore, claims frequency is usually defined as

$$\text{Frequency} = \frac{\text{number of claims}}{\text{time exposure}}, \tag{5.2}$$

where exposure is the amount of time during which the risk has been protected by the insurance company. This is usually measure in years and is called *risk year* (RY). Just to be clear, if a policy is insured for a full year then $RY = 1$, whereas if it is insured for only half a year then $RY = 0.5$.

Let $N(t)$ be the number of claims allocated to a policy during a time interval $[0, t]$, with $N(0) = 0$. The stochastic process $\{N(t) : t \geq 0\}$ is called the claims process and is, according to Beard et al. (1984), a Poisson process if assumptions 1, 2 and 3 are made. Therefore, we are motivated to assume a Poisson distribution (see table 4.1) for the number of claims of an individual policy during any given period of time. Also,

[1] *https://www.towerswatson.com/*

because of the independence of policies, is possible to assume a Poisson distribution at a tariff cell level.

Thus, the first approach to model frequency of claims that always appears in literature is a GLM with Poisson distribution. The link function used is usually the canonical link function, $g_c(\mu) = log(\mu)$, because it keeps the model as multiplicative. However, it is important to remember that, because of the data set structure, not all observations have equal exposure. This means that time must be "weighted" and therefore that an offset must be included in the model.

An offset is a variable used in a Poisson regression to denote the exposure period. If observations are associated with different periods of exposure then, instead of counts, the model must predict a rate, i.e. the number of counts divided by the exposure. In this case, we will have that

$$log(\text{rate}) = \eta \iff log\left(\frac{\text{count}}{\text{exposure}}\right) = \eta \iff log(\text{count}) = \eta + log(\text{exposure}). \quad (5.3)$$

Here, the term log(exposure) is the offset. To put it in simple terms, offset is the log of the time period under study and has a regression coefficient of 1.

5.1.2 Severity Models

Severity is, according to the IRMI[2], *the amount of damage that is (or that may be) inflicted by a loss or catastrophe.* As such, it is essential for an insurance company to study and estimate the severity of claims in order to protect the fund and ensure profit. So, the idea is to build a model that, for each possible risk profile, i.e. tariff cell, predicts the average cost of a claim. Naturally, that means that high severity claims are more expensive than average and low severity claims are less expensive.

Using the information available in the company's dataset, the model has to be capable of forecasting the behavior of new clients based on historical data of clients with similar risk profile. For each tariff cell, (average) severity is calculated as

$$\text{Severity} = \frac{\text{costs of claims}}{\text{number of claims}}, \quad (5.4)$$

where each observed claim has a weight $w = 1$.

[2]International Risk Management Institute, *https://www.irmi.com/online/insurance-glossary.*

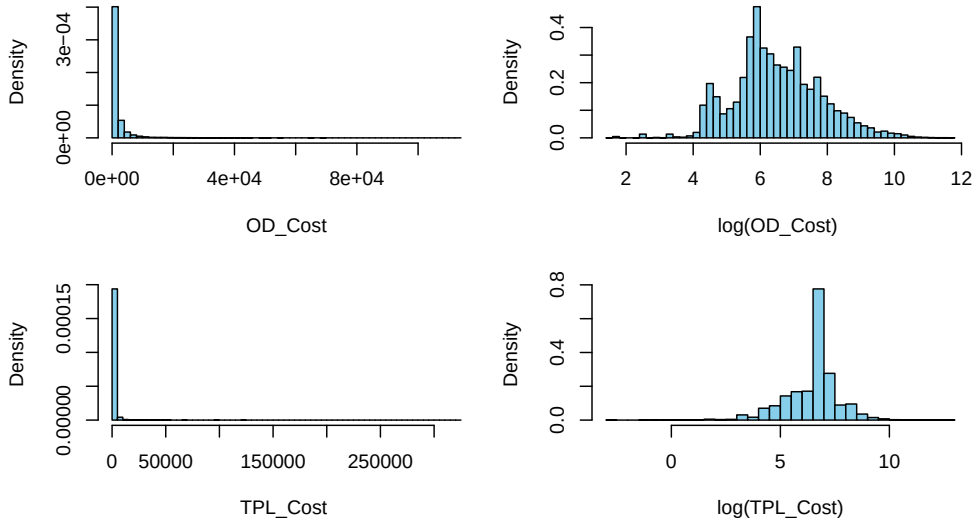


Figure 5.1: Data distribution of cost of claims for Own Damage (OD) and Third Part Liability (TPL).

The distribution of costs of claims is usually positive and skewed to the right and the same happens in our data set, as shown in figure 5.1. This means that the normal distribution, that is often the first distribution suggested, is not suitable. There are, however, several other candidates that fulfill the requirements, but the gamma distribution (see table 4.1) has become the principal (and sometimes the only) distribution used, as we can see in Ohlsson and Johansson (2010), Parodi (2014) and Kaas et al. (2009). One of the reasons why gamma is so appreciated is because it implies that the standard deviation is proportional to the mean, i.e., we have a constant coefficient of variation which seems plausible and more realistic than having a constant standard deviation: say that we have a tariff cell with mean 20 and standard deviation 4, then in another cell with the same exposure but with mean 200 we would rather expect a standard deviation of 40 than of 4 (Ohlsson and Johansson (2010)).

Thereby, claims severity is usually estimated by a GLM with gamma distribution. Also, in order to keep the model as multiplicative, the canonical link function, $g_c(\mu) = -1/\mu$, is usually replaced by the logarithmic function, $g_c(\mu) = log(\mu)$.

5.1.3 The Classical Tariff

Generalized linear models are easily fit in *R* using the **glm()** function of the `stats` library (R Core Team (2016)). This function has the advantage of allowing different combinations of family distributions and link functions. If no link function is specified, the function assumes the canonical link function of the distribution.

We start by creating frequency and severity models according to the classical approach presented in sections 5.1.1 and 5.1.2 respectively. The resulting models **Freq M1** and **Sev M1** are as follows:

- **Freq M1:** Poisson GLM with logarithmic link function

Variable		Estimate	Std. Error	p-value
(Intercept)		-1.774	0.051	<0.001
AgeVehicle[1]		-0.550	0.032	<0.001
AgeDriver[1]		-0.054	0.008	<0.001
HP[2]		0.114	0.023	<0.001
Brand	B1	0.000	-	-
	B2	0.066	0.030	0.028
	B3	0.156	0.031	<0.001
	B4	0.254	0.031	<0.001
Seats	1-3	0.128	0.038	0.001
	4-5	0.000	-	-
	6+	-0.144	0.058	0.013
Use	P	0.000	-	-
	NP	0.618	0.055	<0.001
Fuel	D	0.000	-	-
	G	-0.252	0.026	<0.001
Region	Center	0.101	0.023	<0.001
	North, Lisbon & TV and South & Islands	0.000	-	-
Residual deviance		47808		
AIC		65513		

Table 5.1: Coefficients, standard error and p-values of **Freq M1**. ([1])Fitted variables in decades. ([2])Fitted variable divided by 100.

- **Sev M1:** Gamma GLM with logarithmic link function

Variable		Estimate	Std. Error	p-value
(Intercept)		7.179	0.085	<0.001
AgeDriver[1]		-0.071	0.013	<0.001
HP[2]		0.458	0.049	<0.001
Brand	B1 and B2	0.000	-	-
	B3	0.083	0.044	0.058
	B4	0.280	0.059	<0.001
Use	P	0.000	-	-
	NP	-0.451	0.09507	<0.001
Region	North and Lisbon & TV	0.000	-	-
	Center	-0.140	0.041	0.001
	South & Islands	-0.402	0.102	<0.001
Residual deviance			16168	
AIC			157959	

Table 5.2: Coefficients, standard error and p-values of **Sev M1**. ([1])Fitted variables in decades. ([2])Fitted variable divided by 100.

The models were obtained using the following methods: (i) stepwise selection with the **stepAIC()** function of the `MASS` package (Venables and Ripley (2002)), starting with the respective complete model, in order to achieve the model with best AIC; (ii) hypothesis tests such as Wald's andχ^2 test to decide if variables must be kept in the model; (iii) grouping classes of categorical variables if no differences between them were found. In the end, the model considered is the one with lower AIC where all variables and all classes are significant. Also, categorical nominal variables have as reference the class with the highest exposure, whereas categorical ordinal variables have as reference class the lowest level.

Table 5.1 give us several pieces of informations about the expected frequency of claims obtained with the first model. For example, we can conclude that claims frequency decreases as `AgeVehicle` or `AgeDriver` increases. In fact, for each ten more years, `AgeDriver` decreases claims frequency in 5,3% (since $e^{-0.054} = 0.947$) and `AgeVehicle` decreases frequency of claims in 42.3%. On the other hand, we see that the frequency of claims increases as `HP` increases, at a rate of 12.1% for every 100 `HP`. As for the categorical variables, the ordinal variable `Brand` increases frequency as it goes from lower to higher classes. Also, living in the Center zone or not using a private vehicle

Variables		Frequency relativities	Severity relativities	**Tariff relativities**
AgeVehicle[1]		0.577	1.000	**0.577**
AgeDriver[1]		0.947	0.932	**0.883**
HP[2]		1.120	1.581	**1.771**
Brand	B1	1.000	1.000	**1.000**
	B2	1.068	1.000	**1.068**
	B3	1.168	1.086	**1.269**
	B4	1.290	1.324	**1.707**
Seats	1-3	1.136	1.000	**1.136**
	4-5	1.000	1.000	**1.000**
	6+	0.866	1.000	**0.866**
Use	P	1.000	1.000	**1.000**
	NP	1.855	0.637	**1.181**
Fuel	D	1.000	1.000	**1.000**
	G	0.777	1.000	**0.777**
Region	North	1.000	1.000	**1.000**
	Center	1.106	0.869	**0.962**
	Lisbon & TV	1.000	1.000	**1.000**
	South & Islands	1.000	0.669	**0.669**

Table 5.3: Final tariff relativities for combined models **Freq M1** and **Sev M1**. ([1])Fitted variables in decades. ([2])Fitted variable divided by 100.

both increase claims frequency. As for the variable `Seats`, vehicles with one to three seats are expected to have more claims than those with four or five seats, whereas vehicles with six or more seats are expected to have less.

The interpretation of the severity model, presented in table 5.2 is very similar. Note that these models do not necessarily employ the same variables neither the same joint classes of categorical variables. However, it is interesting to notice that for those variables presented in both models, most coefficients have the same signal in both models which means that, in general, the characteristics associated to policies with higher frequency of claims are also associated with more severe cost of claims. However, there are exceptions, being the most dramatic the variable `Use` - policies with non-private use of the vehicle are expected to have more claims, but claims less severe than the average.

The coefficients in tables 5.1 and 5.2 measure the impact on the response of changing from one class to another or of increasing a unit measure in continuous variables, but

they do not represent the models' relativities. Because both models employ a link function different from the identity, the relativities are obtained only after applying the inverse link function to the linear predictor, as in equation 4.6. Also, because both **Freq M1** and **Sev M1** are multiplicative models, the tariff combined relativities of each variable or variable class are the product of the relativities obtained from each model. These combined relativities are presented in table 5.3 and they reflect the following combined tariff:

$$\begin{aligned}\text{Pure Premium} = \ & \text{Base Term} \times (0.577)^{\text{AgeVehicle}/10} \times (0.883)^{\text{AgeDriver}/10} \times \\ & \times (1.771)^{\text{HP}/100} \times 1.068\ Z_{\text{Brand B2}} \times 1.269\ Z_{\text{Brand B3}} \times \\ & \times 1.707\ Z_{\text{Brand B4}} \times 1.136\ Z_{\text{Seats 1-3}} \times 0.866\ Z_{\text{Seats 6+}} \times \\ & \times 1.181\ Z_{\text{Use NP}} \times 0.777\ Z_{\text{Fuel G}} \times 0.962\ Z_{\text{Region Center}} \times \\ & \times 0.669\ Z_{\text{Region South \& Islands}}\end{aligned} \tag{5.5}$$

where Z_i are dummy variables to indicate classes of categorical variables.

5.2 Potential Problems with Frequency Models

As explained before, overdispersion is probably the greatest challenge with count models. It can be tested using the function **dispersiontest()**, from library `AER` (Kleiber and Zeileis (2008)). This function assesses the following hypothesis:

- $H_0 : \text{Var}(Y) = \mu$;
- $H_1 : \text{Var}(Y) = \mu + \alpha \times \text{trafo}(\mu)$.

The transformation function trafo(μ) is a user specified positive function. Considering trafo(μ) $= \mu$ corresponds to a quasi-Poisson model. In this case, variance can be expressed as

$$\text{Var}(Y) = \mu + \alpha \times \mu = (1 + \alpha) \times \mu = \phi \times \mu. \tag{5.6}$$

By default the latter dispersion formulation is used, testing whether $\phi > 1$. Defining trafo(μ) $= \mu^2$will correspond to a negative binomial (NB) model with quadratic variance function.

The result of the dispersion test applied to model **Freq M1**, the Poisson GLM with link function, is the following:

```
> dispersiontest(FreqM1)

Overdispersion test

data:  m1
z = 7.1979, p-value = 3.058e-13
```

```
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  1.059154
```

Hence, we can reject the null hypothesis that there is no overdispersion with a confidence level of 95%. More over, the function estimates the true dispersion as $\hat{\phi} \approx 1.059$. Therefore, there is evidence to believe that even thought there is overdispersion, there is no need to employ the negative binomial distribution.

Another common problem with count data regressions is excess of zeros. All Poisson, quasi-Poisson and Negative Binomial distributions assume that the count data being modeled have zero counts, but the number of zeros must be carefully examined. Even more, it is not uncommon for overdispersion and excess zeros to go hand in hand, which means that a small overdispersion may even be corrected if a correction for the number of zeros, such as using a zero-inflated model, is considered.

Zero-inflated models can be implemented using the R package `pscl`, Zeileis et al. (2008) and Jackman (2017). This package provides the function **zeroinfl()**, a function with many parameters, three of which deserve special attention:

- `formula`: It is a symbolic description of the model which can be used to specify both components of the model. A formula of type `y ~ x1 + x2` uses the same regressors in both components. If preferable, a different set of regressors could be specified, e.g. `y ~ x1 | z1 + z2`, where the `x1` is being used to predict the conditional count data and `z1 + z2` to predict the inflation of zeros.

- `dist`: Specifies the family to be used in count model, always with a logarithmic link function. The available families are Poisson (`'poisson'`), negative binomial (`'negbin'`) and geometric (`'geometric'`).

- `link`: Specifies the link function to be used in the zero-inflation part of the model, i.e. the binomial model. The available functions are `'logit'`,`'probit'`, `'cloglog'`, `'cauchit'` and `'log'`.

Some fit tests for zero-inflated models have generated divergent opinions within the scientific community regarding their use and appropriateness. The discussion relates to whether a zero-inflated model can be considered nested within another or even comparable (Hilbe (2014), Wilson (2015)). Notwithstanding, the Vuong test for non-nested models will be used during the analysis, keeping in mind the discussion in the assessment of fit.

5.2.1 Model Comparison

In this section, we intend to evaluate the performance of the Poisson model when compared with other counting models. Thus, in order to deal with the initial frequency model overdispersion that the test seems to suggest, the idea is to test not only the adequacy of a negative binomial distribution but also the adequacy of zero-inflated models.

Three models combining previously referred methods were constructed using a few different model selection techniques, such as AIC, BIC, ANOVA and also Wald statistics. As before, in the end we obtained models where all variables are significant. In addition, in the case of categorical variables, all classes that did not present significant differences between each other were grouped in order to obtain variables where all classes are significantly different. In these cases, the reference class was defined as the one with the highest exposure.

So now, the idea is to compare **Freq M1**, a Poisson GLM with link function, with the following models:

- **Freq M2:** Negative Binomial GLM with logarithmic link function;
- **Freq M3:** Zero-inflated Poisson model;
- **Freq M4:** Zero-inflated Negative Binomial model.

The following table provides different measures of goodness-of-fit, providing a basis for deciding which frequency model is the best.

	Number of claims					Total mispredicted values	Sum of Pearson's Residuals	AIC	BIC
	0	1	2	3	≥ 4				
Observed	99509	8025	566	44	6	-	-	-	-
Model Freq M1	99393	8229	500	26	1	**409**	215908	65513	65767
Model Freq M2	99634	7718	711	76	11	614	213987	65353	65629
Model Freq M3	99749	7556	775	65	5	940	182996	**64904**	**65243**
Model Freq M4	99758	7538	782	66	5	975	**181809**	64921	65282

Table 5.4: Comparison of models **Freq M1**, **Freq M2**, **Freq M3** and **Freq M4** by its mispredictions, Pearson's residuals, AIC and BIC.

The package `pscl` (Jackman (2017)) provides a tool for applying the Vuong test, the function **vuong()**. This test is especially useful for comparing zero-inflated count models with non-zero-inflated models. With this function, a positive test statistic provides evidence of the superiority of the first model over the second, while a large, negative test statistic is evidence of the superiority of the second model over the first.

Models Tested			Preferable model	p-value
Freq M1	&	Freq M2	Freq M2	<0.001
Freq M1	&	Freq M3	Freq M3	<0.001
Freq M1	&	Freq M4	Freq M4	<0.001
Freq M2	&	Freq M3	Freq M3	<0.001
Freq M2	&	Freq M4	Freq M4	<0.001
Freq M3	&	Freq M4	Freq M3	0.029

Table 5.5: Vuong tests between the four candidates for better frequency model - **Freq M1**, **Freq M2**, **Freq M3** and **Freq M4**.

Vuong tests seem to prefer zero-inflated models over non-zero-inflated models. Even more, when tested again each other, there is evidence of the ZIP model being superior to the ZINB. Also, the ZIP is the model with lower AIC and BIC, and despite the fact of having a high number of mispredicted values, it is very well behaved in the tail distribution.

So, in conclusion, the model which best performs seems to be **Freq M3**, the ZIP model. The only problem with it is that, since ZIP is a mixture of models, the model multiplicative property is lost, meaning that the tariff can no longer be presented as a base term times relativities. Even though actuaries tend not to enjoy it, it is really not a problem at all - the pure premium can still be easily calculated, although it is not easily presented in a table. Moreover, using zero-inflated models to create a tariff may be an advantage to an insurance company since it provides an additional tool to identify clients who tend not to have or report accidents.

The ZIP model constructed is displayed at table 5.6. The table presents the coefficients obtained for both parts of the model, the logistic regression and the Poisson count process. However, these coefficients have different interpretations. The interpretation of the Poisson regression coefficients is the same as previous. Basically, as

$$log(\mu_i) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n, \tag{5.7}$$

the model is a multiplicative model which means that coefficients may be expressed as relativities,

$$\mu_i = e^{\beta_0} \times e^{\beta_1 X_1} \times e^{\beta_2 X_2} \times ... \times e^{\beta_n X_n}. \tag{5.8}$$

As for the logistic regression, the interpretation of parameters is slightly different. Because the link function used is the logit function, the relation between the probability of success (in our case the probability of zero) and the linear predictor is

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta \iff \pi_i = \frac{1}{1+e^{-\eta}}. \tag{5.9}$$

Notice that the ratio $\frac{\pi_i}{1-\pi_i}$ is the odds ratio (OR) of success against not success. Suppose now, that we want to analyze the coefficient of a certain variable, for example the coefficient β_1 of a dummy variable X_1, which is coded as 0 or 1. Then, we have

$$X_1 = 0 \Longrightarrow \frac{\pi(X_1 = 0)}{1 - \pi(X_1 = 0)} = e^{\beta_0 + \beta_2 X_2 + \ldots + \beta_n X_n} , \tag{5.10}$$

$$X_1 = 1 \Longrightarrow \frac{\pi(X_1 = 1)}{1 - \pi(X_1 = 1)} = e^{\beta_0 + \beta_1 + \beta_2 X_2 + \ldots + \beta_n X_n} . \tag{5.11}$$

Hence

$$OR(X_1 = 0, X_1 = 1) = \frac{\frac{\pi(X_1=1)}{1-\pi(X_1=1)}}{\frac{\pi(X_1=0)}{1-\pi(X_1=0)}} = \frac{e^{\beta_0 + \beta_1 + \beta_2 X_2 + \ldots + \beta_n X_n}}{e^{\beta_0 + \beta_2 X_2 + \ldots + \beta_n X_n}} = e^{\beta_1}. \tag{5.12}$$

In conclusion, the exponential of a coefficient gives us the odds ratio of changing from the reference class to the class of the variable's coefficient, if all other variables are kept constant.

Variables		Zero-inflation model coefficients (binomial with logit link):			Count model coefficients (poisson with log link):		
		Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
(Intercept)		-2.926	0.196	<0.001	-1.980	0.069	<0.001
AgeVehicle[1]		1.126	0.069	<0.001	-	-	-
AgeDriver[1]		0.125	0.017	<0.001	-	-	-
HP[2]		0.436	0.079	<0.001	0.325	0.048	<0.001
Brand	B1	-	-	-	0.000	-	-
	B2	-	-	-	0.062	0.031	0.047
	B3	-	-	-	0.151	0.032	<0.001
	B4	-	-	-	0.225	0.033	<0.001
Seats	1-3	0.742	0.154	<0.001	0.459	0.088	<0.001
	4-5	0.000	-	-	0.000	-	-
	6+	0.425	0.116	<0.001	0.000	-	-
Use	P	-	-	-	0.000	-	-
	NP	-	-	-	0.603	0.059	<0.001
Fuel	D	0.000	-	-	-	-	-
	G	0.487	0.053	<0.001	-	-	-
Region	North	-	-	-	0.000	-	-
	Center	-	-	-	0.089	0.025	<0.001
	Lisbon & TV	-	-	-	0.000	-	-
	South & Islands	-	-	-	0.000	-	-

Table 5.6: Coefficients, standard error and p-values of **Freq M3**, a ZIP model with two components, a logistic model that generates structural zeros (on the left) and a Poisson count model (on the right). ([1])Fitted variables in decades. ([2])Fitted variable divided by 100.

It is interesting to observe the signals of the obtained coefficients. In the zero-inflation model, all coefficients are positive meaning that increasing a unit in continuous variables or changing from the reference class to other classes increases the probability of generating a zero, therefore decreasing the resulting frequency. On the other hand, on the Poisson count model, increasing a unit in continuous variables or changing from the reference class to other classes increases the resulting frequency of claims.

As explained before, because not all data set observations have equal exposure, and because we are modeling a frequency and not the raw number of claims, a model offset must be considered. However it must only be included in the Poisson count model and not in the logistic regression - as it is only predicting zeros or non-zeros, not a rate. Thus, to adjust the logistic regression to the observations' different exposure, the variable RY was added to the model as a covariate, even though the coefficient obtained is not relevant for inclusion in the tariff. The inclusion of this variable revealed to be significant in the model (p-value < 0.001).

5.3 The Ideal Severity Model Distribution

As explained in section 5.1.2, the distribution of the cost of claims is usually a heavy-tailed distribution. In fact, for some insurance portfolios, a substantial part of the total claim cost may be due to a few large claims. Such dominating claims can make estimates very volatile and their effect has to be reduced somehow (Ohlsson and Johansson (2010)). In order to deal with these large claims, it is usual to truncate the distribution of claims, leaving out claims above a defined value c. The choice of the threshold c is very subjective, depending on the distribution and the analyst. If in one hand c must be chosen as large as possible in order to cover as many claims as possible, on the other hand it has to be chosen small enough to give reliable estimates. In our case, based on claims severity quantile distribution presented in figure 5.2, the threshold was defined as the quantile 99. Table 5.7 presents the cumulated cost of claims and the cumulated number of claims on the upper tail distribution, by quantiles. Hence, cutting the distribution on the 99th percentile, we are excluding less than less than 100 claims which together represent more than 7% of the total amount of claims in our data set.

	$Q_{0.90}$	$Q_{0.91}$	$Q_{0.92}$	$Q_{0.93}$	$Q_{0.94}$	$Q_{0.95}$	$Q_{0.96}$	$Q_{0.97}$	$Q_{0.98}$	$Q_{0.99}$
Cumulated cost of claims	42.09%	39.48%	36.68%	33.60%	30.28%	26.68%	22.66%	18.15%	13.06%	7.17%
Cumulated number of claims	953	858	763	667	572	477	382	286	191	96

Table 5.7: Cumulated cost of claims and number of claims on the upper tail distribution, by distribution quantiles.

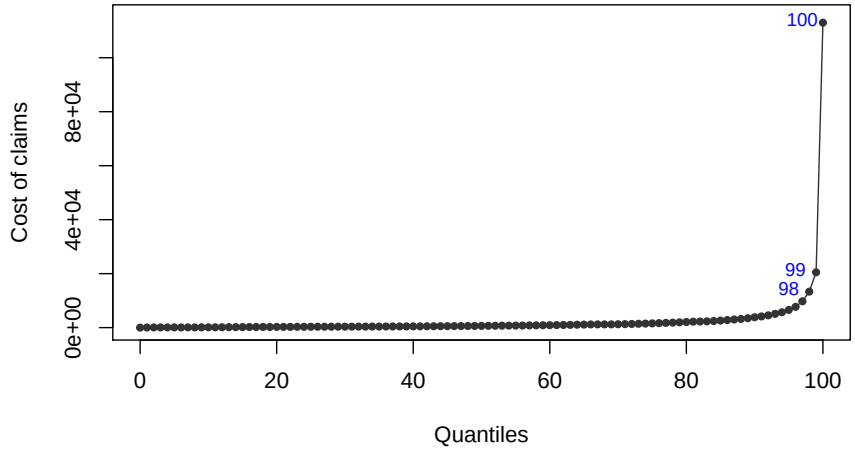


Figure 5.2: Quantile-plot of severity of claims empirical distribution.

Sometimes, in more detailed analysis than the one we are presenting, large claims are modeled separately. They can be especially important to model for reinsurance analysis because this kind of claims is usually protected by reinsurance. The analysis of large claims separately will not be explored here, but some good references for this topic are Boelviken (2014) and Ohlsson and Johansson (2010).

Even after excluding large claims from our data set, the distribution of claims still has a heavy tail. For that reason, the next step should be deciding which known distribution approximates better our empirical distribution. By default, actuaries always use a Gamma distribution, as defined in the classical approach. But, is there any other distribution that fits better this kind of data? To answer this question, the following distributions were considered:

- **Exponential distribution**: an exponential family distribution defined in table 4.1.

- **Lognormal distribution:** A continuous variable X follows a Lognormal distribution of parameters $(\mu, \sigma) \in \mathbb{R}^2$, $\sigma > 0$, $X \sim \text{Lognormal}(\mu, \sigma)$, if its pdf is defined as

$$f(x|\ \mu, \sigma) = \frac{1}{\sqrt{2\pi}x\sigma} e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}} \tag{5.13}$$

  where μ is the mean and σ the standard deviation.

- **Gamma Inverse distribution:** A positive continuous variable X follows a Gamma Inverse distribution of parameters $(\alpha, \beta) \in \mathbb{R}^2_+$, $X \sim \text{GI}(\alpha, \beta)$, if its pdf is defined as

$$f(x|\ \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}} \tag{5.14}$$

  where α is the shape parameter and β the scale parameter.

- **Weibull distribution:** A positive continuous variable X follows a Weibull distribution of parameters $(\gamma, \beta) \in \mathbb{R}_+^2$, $X \sim \text{Weibull}(\gamma, \beta)$, if its pdf is defined as

$$f(x|\ \gamma, \beta) = \frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1} e^{-\left(\frac{x}{\beta}\right)^{\gamma}} \tag{5.15}$$

  where β is the scale parameter and γ the shape parameter.

- **Pareto distribution:** A positive continuous variable X follows a Pareto distribution of parameters $(\gamma, \beta) \in \mathbb{R}_+^2$, $X \sim \text{Pareto}(\gamma, \beta)$, if its pdf is defined as

$$f(x|\ \gamma, \beta) = \frac{\gamma \times \beta^{\gamma}}{(x+\beta)^{\gamma+1}} \tag{5.16}$$

  where β is the scale parameter and γ the shape parameter.

In order to compare the adequacy of these distributions to our data, the five distributions were fitted, as well as the Gamma distribution. The function used to fit the distributions was the function **fitdist()** from package `fitdistrplus`, Delignette-Muller and Dutang (2015).

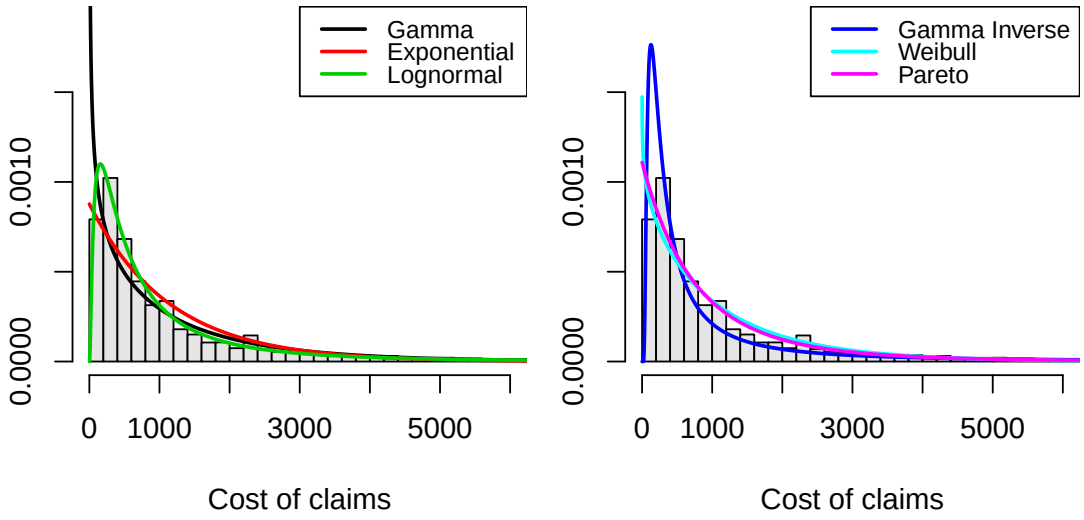


Figure 5.3: Histogram of cost of claims and some fitted distributions.

Figures 5.3 and 5.4 give us a visual idea of how good these distributions approximate the data. In particular, from figure 5.4 we verify that overall the Gamma distribution does not satisfactorily fit costs of claims because the pattern does not follow a straight line. Based on these plots, we can conclude that the Lognormal and the Pareto distributions seem to be adequate to approximate cost of claims distribution.

In order to have a more precise measurements to decide which distribution is preferable, the average quantile absolute error (AQAE) was calculated for each distribution, i.e. the average of the absolute difference between the quantiles of the empirical

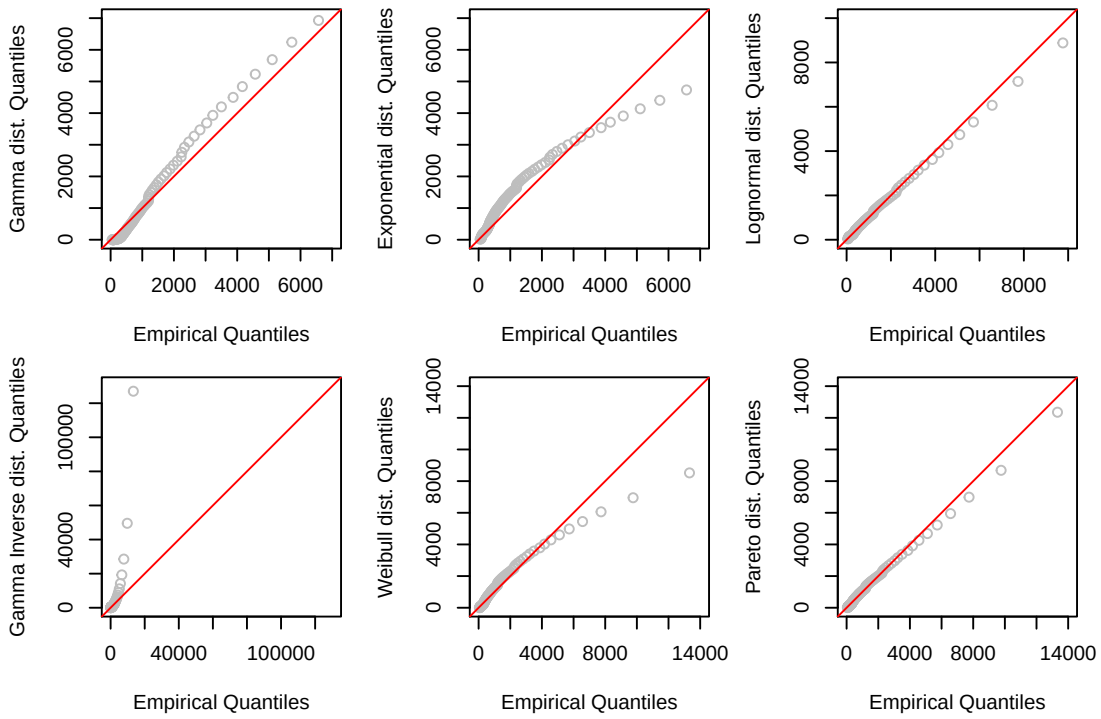


Figure 5.4: Quantile-Quantile plot for comparison empirical cost of claims distribution with some fitted distributions, without large claims.

distribution and each distribution estimated quantiles. The results, presented in table 5.8, indicate the best distributions among those tested are the Lognormal, the Pareto and the Gamma distributions, in that order.

Distribution	AQAE
Gamma	254
Exponential	432
Lognormal	71
Gamma Inverse	2295
Weibull	267
Pareto	100

Table 5.8: Average quantile absolute error (AQAE) of some fitted distributions to cost of claims data, without large claims.

So, ideally, cost of claims should be modeled considering that it follows a Lognormal distribution. The problem is that this distribution is not a member of the exponential family and therefore cannot be fitted as a GLM. When a response variable is believed to follow a Lognormal distribution, the usual way of fitting a regression model is to log-transform the response variable, making it a normally distributed variable, and

then apply a linear regression. However, because

$$E(log(Y)) \neq log(E(Y)), \tag{5.17}$$

fitting the log-transformation of the response through a linear regression will not result in the same model we would obtain if we could fit a GLM with a Lognormal distribution. As for the Pareto distribution, since it is not related in a simple way with the exponential family, it is not compatible with a GLM.

All other distributions showed to be less adequate for fitting cost of claims than Gamma. In particular, the Exponential distribution, the only tested distribution besides Gamma which belongs to the exponential family, performed poorly in this analysis, achieving only a better AQAE than Gamma Inverse distribution.

The lack of alternatives to deal with cost of claims leads us to explore the log-transformation of data, despite the inequality 5.17. Besides fitting a linear regression to the transformed data as explained before, a Gamma GLM was also considered by suggestion of Boelviken (2014). As before, in order to discuss which distribution better fits the transformed data, we compared the quantiles of both distributions with the empirical quantiles. The results, presented in figure 5.5 and in table 5.9, suggest that the Gamma distribution is more adequate than the Normal distribution.

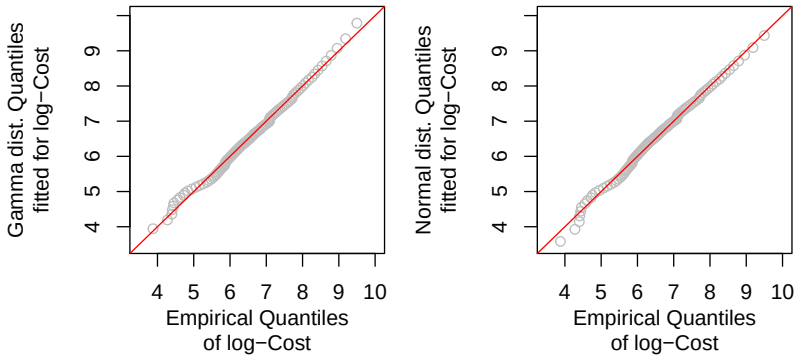


Figure 5.5: Quantile-Quantile plot for comparison empirical cost of claims distribution with Gamma and Normal distribution, after log-tranform cost of claims, without large claims.

Distribution (after log-transformation)	AQAE
Gamma	0.0608
Normal	0.0613

Table 5.9: Average quantile absolute error (AQAE) of Gamma and Normal distributions fitted to log-transformed cost of claims, without large claims.

5.3.1 Some Conclusions About the Severity Model

We have seen that the Gamma distribution is not the ideal distribution to represent cost of claims. The problem here is that all other preferable distributions are not members of the exponential. Therefore, we concluded that the only other viable option is to fit a Gamma GLM for the log-transformed data despite the fact that the expected cost of claims predicted by the model might be wrong. Having this in mind, is it really preferable to use data transformation, or are the differences between the Gamma GLM model and the Gamma GLM with transformed data are small enough not to consider the latter?

In order to answer this question, we used methods such as AIC criteria, Wald's test and grouping classes of categorical variables to obtain the best possible final models. Again, in these final models, all variables and classes considered are significant. The data set used to fit the models was cut at the 99th percentile. The obtained model were **Sev M2** and **Sev M3**:

- **Sev M2**: Gamma GLM (without large claims)
- **Sev M3**: Gamma GLM with log-transformation of costs (without large claims)

	Min	$Q_{0.25}$	$Q_{0.50}$	$Q_{0.75}$	Max	Correlation with observed costs
Observed costs	5	308	635	1510	20390	-
Predicted costs: Sev M2	665	1309	1447	1601	4118	0.112
Predicted costs: Sev M3	503	635	667	703	1149	0.062

Table 5.10: Summary of observed costs and predicted costs by models **Sev M2** and **Sev M3**.

As we can see from figure 5.6, there is nothing which may indicate that these models are poorly constructed. However, both table 5.10 and figure 5.7 lead us to the conclusion that Gamma GLM seems not to be capturing the variability of data, which compromises the correlation between observed and predicted costs. Even more, the Gamma GLM with log-transformation seems to be even worst, predicting values in a smaller and lower interval and having a lower correlation with observed cost of claims. Based on this, we conclude that model **Sev M2** is preferable to model **Sev M3**, i.e. is preferable to use the untransformed data rather than log-transformed.

An alternative hypothesis to a Gamma GLM is a pareto regression, proposed by Beirlant and Goegebeur (2003). This method is based on the transformation of the dependent variables into generalized residuals and on an exponential regression model for

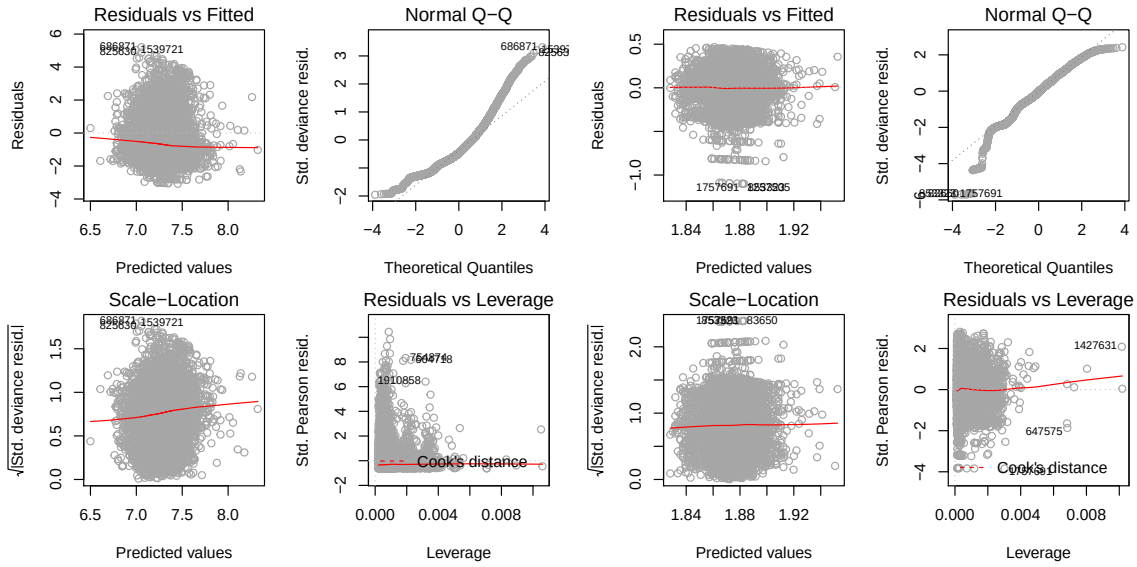


Figure 5.6: Model output plots - models **Sev M2** (the four plots on the left) and **Sev M3** (four plots on the right).

these residuals, with parameters being estimated by the maximum likelihood method. However, according to authors, in practice this method does not always work, since the models have problems finding out an appropriate relationship between the response and the covariate information and therefore finding out the ideal transformation.

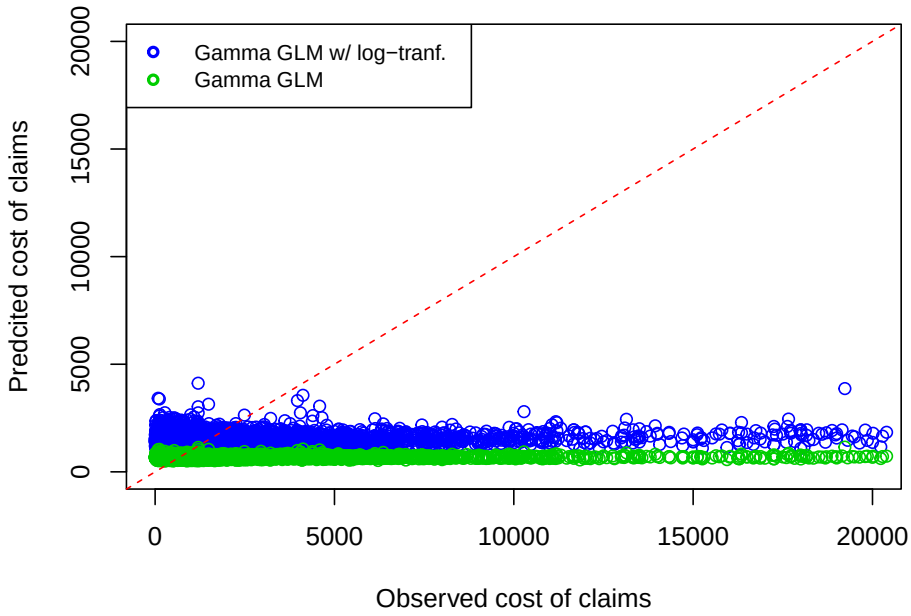


Figure 5.7: Predicted cost of claims by models **Sev M2** and **Sev M3** against observed costs.

5.4 Smooth Functions

As previously explained, the idea behind smooth a data is to create an approximating function that attempts to capture important patterns while leaving out noise. They are usually used so simplify continuous variables, but they can also be used with ordinal categorical data.

Smooth functions are introduced in regressions trough Generalized Additive Models (GAM), which can be implemented in *R* using the package `mgcv`. This package, developed by Simon Wood, includes a wide variety of smoothers and distributions beyond the exponential family, as well as a **gam()** function which will be the base of analysis used from this point on. Some good literature to support the methods used in this package are Wood (2006), Wood (2011), Wood et al. (2016), Wood (2004), and Wood (2003).

5.4.1 Time Effect

Taking into account that our data set has a period of exposure of six years, some kind of time effect might be causing noise or distortion on frequency or severity models. For this reason, it is crucial in this kind of analysis to study the time effect on models and, if any significant result is achieved, to make an appropriate adjustment.

There are two different potential time effect on this kind of data. The first one is the seasonal effect throughout the year, i.e. the effect of the variable `Month`. Because policies are usually under risk for a period of one year, we expect all months to have a similar exposure to risk. If no seasonal effect is detected, we would expect all months to have similar frequency of claims with similar severities, meaning that weather or other consequences of seasons such as holidays do not have impact on our data. On the other hand, it is also possible to detect differences in the response variables as years pass. Even though six years may not be enough to detect behavioral differences in societies, it is enough to detect things like inflation or economy depressions. For that reason, the effect of the variable `Year` must also be studied.

Time effect on Severity

The simplicity of the severity data set makes it easy to study the time effects of variables `Month` and `Year` - each claim is allocated to a day of occurrence, i.e. to a unique month and year. Therefore, in order to test the effect of these variables, we fitted a Gamma GAM model with cubic splines for both variables. Since the effect of the variable month is cyclic every 12 months, the variable `Month` was smoothed using a cycling cubic spline, whereas variable `Year` was smoothed with a simple cubic spline.

The smoothed time effect of variables `Month` and `Year` in severity model can be observed in figure 5.8. Interpreting the figure, we conclude that the severity seems to be declining through the years and that claims seem to be more severe than average in September, October and November, the first months of rain after summer and the begging of the school year, and less severe during the summer. However, not both variables influence significantly the response variable. In fact, when included in a Gamma GAM model, the effect of variable `Month` is significant (p-value < 0.001) and responsible for explaining 0.54% of model's deviance, but the effect of `Year` is not (p-value $= 0.240$). Thus, we conclude that the severity model must be adjusted to seasonal effect.

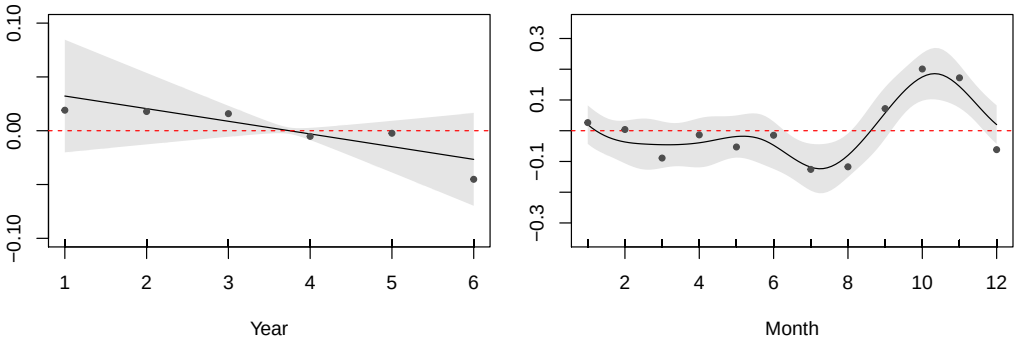


Figure 5.8: Gross Effect of variables `Year` and `Month` in severity of claims.

Time effect on Frequency

Each data set observation is associated with a period of exposure and to a date of beginning of exposure and end of exposure. This period can involve several months but never more than one civil year. This happens because every time the year changes, a new observation referent to the same policy is created. Therefore, the period of exposure may vary from 1 to 365 or 366 days .

The problem of this construction is that each observation, that contains a period of exposure (RY) and a dichotomous variable that indicates if there was or not a claim in that period, cannot be associated to a unique month. Therefore, the seasonal effect cannot be studied in frequency model, concluding that the only possible time effect adjustment is for variable `Year`.

The effect of `Year` in claims frequency is presented in figure 5.9. The effect of this variable is significant (p-value < 0.001) and responsible for explaining 0.10% of model's deviance. This effect shows us that during the third and fourth year the number of reported claims was lower than average, which may be related to an economic depression and a misreport of claims. Therefore, we conclude that the frequency model must be adjusted to the variable `Year`.

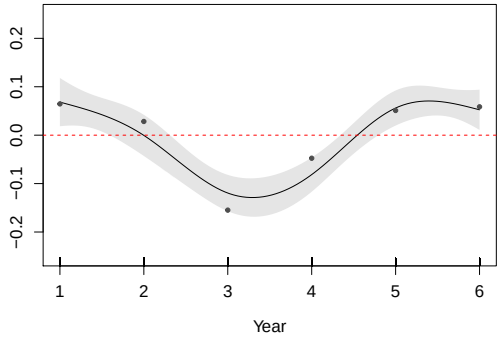


Figure 5.9: Gross Effect of variable `Year` in frequency of claims.

5.4.2 Smoothing Continuous Variables

Smooth functions can also be applied to continuous covariates of a model. If no smooth function is considered, each continuous variable is estimated having a fixed relation with the response, depending on the link function chosen. The advantage of using smooth functions is capturing tendencies which may suffer high variations trough the covariate domain, ensuring however that two close values have close relativities thanks to its method for penalize *wiggliness*.

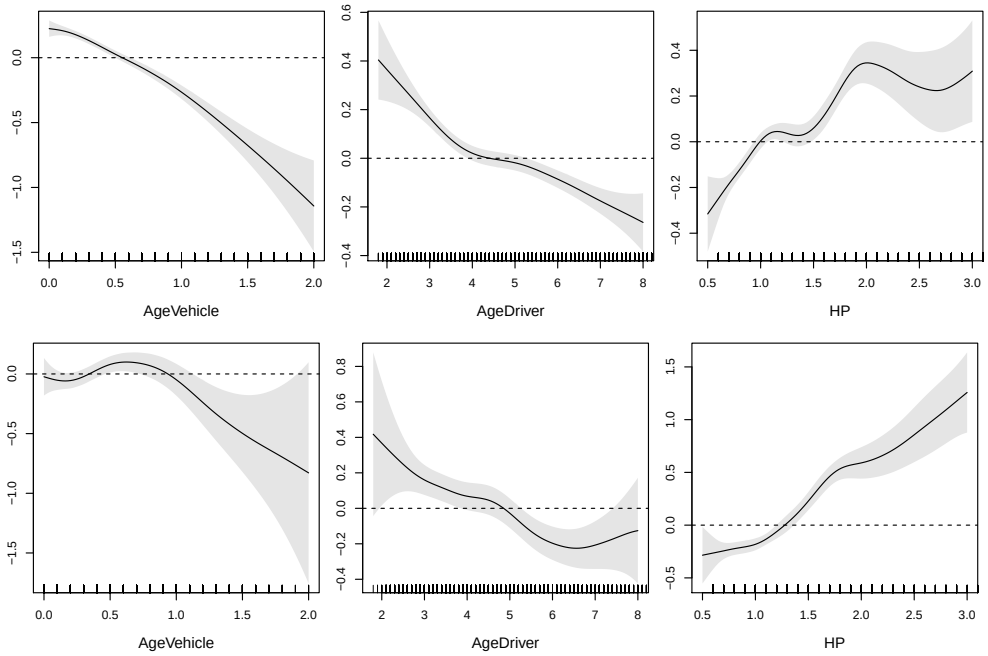


Figure 5.10: Tendency of variables `AgeVehicle`, `AgeDriver` and `HP` captured with smooth splines, for the frequency model (top) and the severity model (bottom).

5.4.3 Obstacles: Zero-Inflated Models in `mgcv`.

Using smooth functions, or more specifically, using a GAM with package `mgcv`, introduces a new obstacle in what has been developed in this dissertation until now on claims frequency. At this point, we have already concluded that the best model to capture effects on claims frequency is a Zero-Inflated Poisson model. Hence, the idea now would be to expand this model to a GAM, introducing smooth functions and use them to represent continuous covariates and adjust the model to the effect of years.

However, not everything is so easy as it seems to be. In fact, package `mgcv` has a family model for applying Zero-Inflated Poisson models: the **ziplss()** function. The problem is that the results obtained using this function are very different from those obtained using the function **zeroinfl()** from package `pscl` and also from a simple **glm()** function.

Variables		Poisson GLM with link function. R function: **glm()**			Zero-inflated Poisson model. R function: **zeroinfl()**			Zero-inflated Poisson model. R function: **gam()** with family **ziplss()**		
		Estimate	Std. Error	p-value	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
Count model										
(Intercept)		-2.438	0.033	<0.001	-1.908	0.049	<0.001	-1.885	0.128	<0.001
HP		0.153	0.021	<0.001	0.158	0.023	<0.001	0.143	0.088	0.104
Brand	B2	0.072	0.030	0.0161	0.068	0.031	0.0277	-0.203	0.113	0.072
Brand	B3	0.155	0.030	<0.001	0.152	0.032	<0.001	0.004	0.109	0.974
Brand	B4	0.300	0.030	<0.001	0.299	0.032	<0.001	0.055	0.113	0.629
Seats	1-3	0.150	0.038	<0.001	0.158	0.039	<0.001	0.289	0.134	0.031
Use	NP	0.492	0.055	<0.001	0.503	0.058	<0.001	0.84	0.145	<0.001
Region	Center	0.129	0.023	<0.001	0.129	0.024	<0.001	0.117	0.084	0.168
Logistic model										
(Intercept)		-	-	-	-0.367	0.086	<0.001	-2.486	0.011	<0.001

Table 5.11: Comparison of two ZIP models estimated using functions **zeroinfl()** and **ziplss()** with their analog GLM model.

Table 5.11 compares a usual GLM with two Zero-Inflated Poisson model with only the intercept to predict the inflation of zeros - one using the function **zeroinfl()** and the other **ziplss()**. When looking at the estimates obtained with the **zeroinfl()** ZIP, we quickly conclude that the coefficients behave similarly, having the same signal and equivalent significance, but the same does not happen for the estimates of the **ziplss()** function. With this function, most of the coefficients are no longer significant and some are even estimated with opposite signals. The reason why this happens is that the models do not use the same methods, as the former uses maximum likelihood and the latter restricted maximum likelihood. According to Corbeil and Searle (1976), this second method is a particular form of maximum likelihood estimation which does not base estimates on a maximum likelihood fit of all the information, but instead uses a

likelihood function calculated from a transformed set of data.

Although both methods are valid, the observed differences make us apprehensive about the latter method. We conclude that it could only be included in the analysis if studied in more detail, after making sure that such dramatic mistakes would be prevented. Therefore, the analysis of claims frequency proceeds using a quasi-Poisson GAM.

5.5 The Final Tariff

After studying each model in detail, we are now in the position to build the final tariff. So that this tariff may be compared with the classical approach, a training data set is used to create the models and a test data set is used to compare them, in a proportion of 70/30.

This section is divided into two subsections. In the first subsection, the methodology developed for combining the different methods into a final tariff is explained and the final models created. Then, in the second subsection, the two tariffs are explored and compared.

5.5.1 Methodology

Combining the methods explored in this dissertation was not an easy task. Ideally, we would like to create a tariff combining data imputation, smooth functions and appropriate models and distributions to predict each response variable. The truth is that combining these methods is computationally challenging, and even though most methods are implemented in R, some are not compatible with each other. Thus, it is important to state here that the presented tariff in this chapter is not the ideal one, but the best achieved considering the available resources. The methodology used, common to frequency and severity models, is described next.

As explained before, the models are created based on each model training data set. As in any insurance data set, the number of missings is a problem and considering only the complete cases might drastically reduce the dimensions of the data set. Thus, the first task is to apply multiple imputation to the training data set, originating m imputed data sets. Secondly, based on the complete cases of the training data set, a base model is constructed. This base model, which is a GAM with smooth functions for continuous variables, adjusted for time effects and that has into consideration all the problems studied before, is used to capture the variables of potential interest for predicting the data.

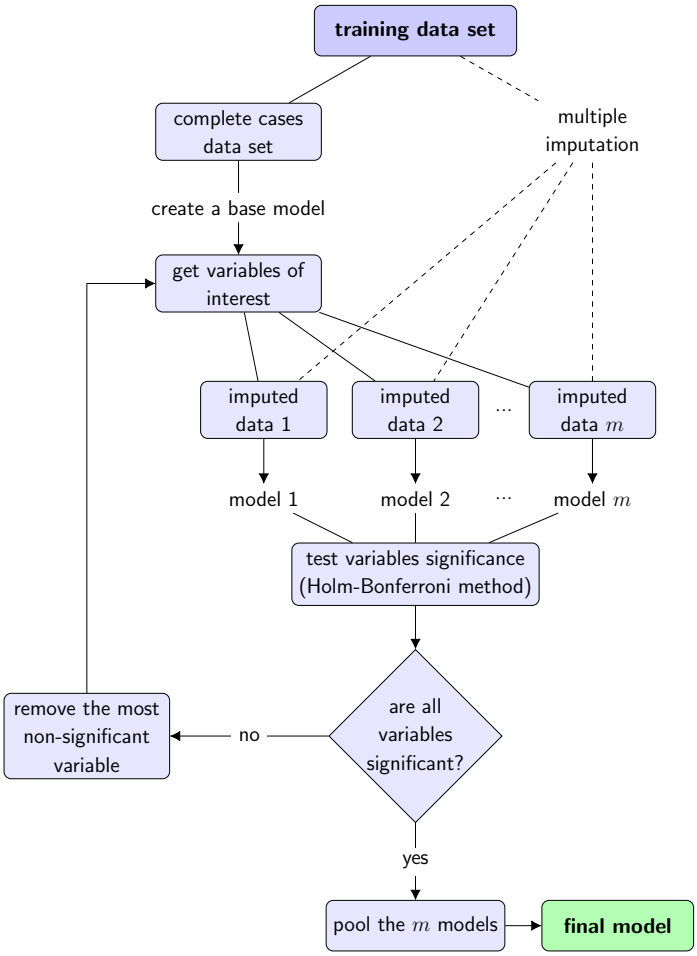


Figure 5.11: Procedure flowchart.

Once the base model is created, the same model is applied to each one of the m imputed data sets, and the results analyzed. Note that the m models have the same structure but different estimated coefficients with different significance levels. In order to test if a variable is overall significant, the **Holm-Bonferroni method** (Holm (1979)) is used. This method, used to counter-attack the problem of multiple comparisons, adjusts the rejection criteria of each of the individual hypotheses. It is a more relaxed variation of the Bonferroni correction (Dunn (1961)), a method known to be very conservative, and is formulated as follows.

Definition 5.5.1 (Holm-Bonferroni method)**.** Let $H_1, ..., H_m$ be a family of hypotheses and $P_1, ..., P_m$ its p-values. Without loss of generality, consider that these p-values (and the respective hypotheses) are ordered from lowest to highest. For a given significance level α, let k be the minimal index such that

$$P_k > \frac{\alpha}{m + 1 - k}. \tag{5.18}$$

Then, the null hypothesis is rejected for all $H_1, ..., H_{k-1}$. In particular, if $k = 1$, none of the hypotheses can be rejected and if no such k exists then all null hypotheses are rejected. □

Thus, a variable is defined to be significant for imputed data sets if all null hypotheses for that variable are rejected according to the Holm-Bonferroni method. If however at least one of the hypothesis associated with a certain variable cannot be rejected, the variable is considered as non-significant and is removed from the variables of interest. Then, the updated model is applied to each one of the m imputed data sets and the process is repeated. Finally, when all variables are considered significant to the imputed data sets, the m models are pooled and a final model is obtained.

Based on this procedure, the final frequency and severity models were created. Together, they estimate each risk's pure premium. In our case, as explained previously, the average λ indicates that $m = 10$ is reasonable number of imputations.

5.5.2 Comparison of Tariffs

As explained before, the final tariff is an attempt to improve in two directions what is currently practiced in the market. First, by dealing with missing values so that the final models are not influenced by a possible bias of the data. Second, by improving models using GAM instead of GLM, so that smoothing functions can be used in order to capture non-linear trends in data. Therefore, the adequacy of the final tariff must be determined in a two-step analysis, deciding first if using a GAM instead of a GLM improves the capacity of predicting the data and then deciding if multiple imputation produces, or not, different results.

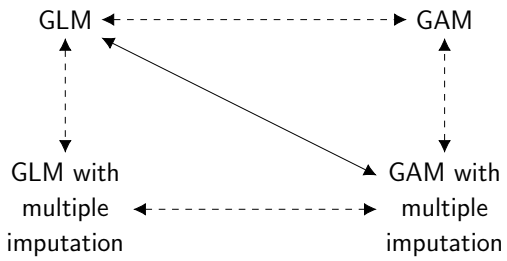


Figure 5.12: Final tariff's analysis diagram.

In order to reach a conclusion about the final tariff, we will now proceed with the four analysis represented with dashed lines on figure 5.12. Based on this figure, it is easy to see that horizontal arrows compare the usage of GLM against GAM, whereas vertical arrows study the adequacy of using multiple imputation.

Starting with the horizontal comparisons, the table 5.12 contains some information about the models and their adjustment to the test data sets. Notice that the models performed without multiple imputation were tested on a complete cases test data set, i.e. considering only the observations without missing data of the test data set. Furthermore, the models constructed using multiple imputation were tested on the imputed test data set, i.e. the test data set imputed using multiple imputation for $m = 1$. In general, it is observed that the GAM models present better fit to the test data than the GLM models, presenting higher explained deviation and higher Pearson's pseudo R^2. In particular, in the severity models, where it is possible to calculate AIC and BIC, GAM models usually have lower values, failing only for models without imputation, where, GLM BIC is lower than the GAM's. However, this criterion heavily penalizes the number of parameters for that reason a result like this was expected, meaning that it does not contradict all other results.

	GLM	GAM	GLM w/ MI	GAM w/ MI
Frequency Model				
Parameters	9	68	9	68
Explained Deviance	1.84%	2.01%	2.26%	2.41%
Pearson's Pseudo R^2	0.0221	0.0165	0.0316	0.3305
Average Predicted Frequency	0.1275	0.1274	0.1495	0.1494
Severity Model				
Parameters	6	56	6	56
Explained Deviance	6.70%	8.08%	7.27%	8.24%
AIC	164186	164074	270632	270454
BIC	164243	164282	270693	270646
Pearson's Pseudo R^2	0.2162	0.2235	0.2845	0.3031
Average Predicted Severity	1702	1671	1661	1633
Pure Premium				
Average Predicted PP	217	213	248	244
Relative Absolute Error	0.8091	0.8101	0.8150	0.8138

Table 5.12: Comparison between GLM and GAM for the final tariff.

As for the vertical comparisons, it is interesting to analyze first the differences observed in the response variables for both data sets, complete cases and imputed. Table 5.13 summarizes this information, making it possible for us to observe that the imputed data set presents a higher claims frequency, although with a lower average severity, resulting in a higher average pure premium. Thus, we can conclude that the set of observations we are excluding when considering only the complete cases have above average frequency of claims, even though with below average severity.

From table 5.14, we can draw some more conclusions about the use of multiple imputation in the context of this dissertation. It is observed that in the complete cases

	Complete Cases Data Set	Imputed Data Set
Observed Frequency	0.1296	0.1507
Observed Severity	1813	1726
Observed PP	235	260

Table 5.13: Observed Frequency and Severity in test data sets.

data set, the models generated without multiple imputation have higher Pearson's pseudo R^2. On the other hand, when tested in the imputed data set, the models with multiple imputation have higher values of Pearson's pseudo R^2. This conclusion is in line with what was expected, meaning that a model constructed from a biased data set will perform better when tested on a biased data set and worst when tested on an unbiased data set, and vice versa. This conclusion also strengthens the idea that a tariff constructed based on a data set where a set of observations were excluded, which may cause a bias in the data, might not correctly estimate a real market where there is no "exclusion" of customers. Also, by the Relative Absolute error, we conclude that the models with multiple imputation present a better fit to the data in both bases.

	Complete Cases Data Set				Imputed Data Set			
	GLM	GLM w/ MI	GAM	GAM w/ MI	GLM	GLM w/ MI	GAM	GAM w/ MI
Frequency Model								
Pearson's Pseudo R^2	0.0221	0.0193	0.0165	0.0116	0.0223	0.0316	0.0269	0.3305
Average Predicted Frequency	0.1275	0.1269	0.1274	0.1267	0.1604	0.1495	0.1611	0.1494
Severity Model								
Pearson's Pseudo R^2	0.2162	0.2081	0.2235	0.2176	0.2542	0.2845	0.2798	0.3031
Average Predicted Severity	1702	1704	1671	1673	1604	1661	1578	1633
Pure Premium								
Average Predicted PP	217	216	213	212	257	248	254	244
Relative Absolute Error	0.8091	0.8084	0.8101	0.807	0.8168	0.8150	0.8197	0.8138

Table 5.14: Comparison between models with and without multiple imputation.

The observed and estimated mean of the response variables also allows an interesting analysis. When we compare the table 5.13 with the table 5.14, we could easily be wrongly led to conclude that all models were under-estimating the required amount to cover expenses associated with claims in our data set. However, this measure of comparison is not highly indicated in our case, because of the heavily-tailed distributions considered. In such cases, the mean of the distribution is always far to the right of the median. Thus, predicting a lower average might simply mean that the model predicts the core data better, shortening the distance between the mean and the median.

Chapter 6

Conclusions

In this chapter, we summarize the main conclusions of the work carried out in this dissertation, as well as some possible directions for future work.

6.1 Main conclusions

This dissertation studied the model and methods behind the creation of a motor insurance tariff. The aim of this work was to present the concept of a tariff, to understand the theory behind its classical formulation and to study ways of improving it. We saw that creating a tariff means defining a model that estimates the pure premium of a policy, which is obtained through the product of two models - a model for the frequency of claims and another for the severity of claims.

The methodologies into practice nowadays on insurance markets are based on Generalized Linear Models (GLM), using a Poisson GLM for frequency and a Gamma GLM for severity, both with logarithmic link functions in order to keep the models as multiplicatives. In particular, these models are often created by insurance software developed only for this purpose, software usually very user-friendly even though not flexible at all, precluding actuaries from exploring other techniques and models.

As such, it was our desire to present new methods for improving the final tariff. The idea was to explore them and describe the creation of a tariff step-by-step. We started by presenting a new idea for how to deal with missing data, preventing data to be excluded from the analysis. We then introduced single imputation although after realizing that it might not be as useful as thought at first, multiple imputation was presented as an alternative. We showed that multiple imputation is a good alternative to complete cases analysis. In fact, even for variables with a high percentage of missing values (around 40%), this method keeps the original data tendency. Thus,

the application of this method makes it possible to fit models with more observations, avoiding the exclusion of observations and its possible distortion of reality.

The classical approach for the frequency model was then considered and a Poisson GLM with logarithmic link function was fitted to our data set. Even though this model presented a reasonable fit, the data was overdispersed and therefore a correction had to be taken into account. As an alternative to correct the overdispersion, a Negative Binomial distribution and/or a zero-inflated model was considered. We concluded that, for claims frequency, the model that best represents the data was a zero-inflated Poisson model. However, practical incompatibilities between R packages led us to abandon this model and choosing the second best one instead, using afterwards the second-best model, a quasi-Poisson GLM with log link.

As for severity models, we started by observing the distribution of costs, realizing that the distribution was very asymmetrical, skewed and heavy-tailed, which led us to consider as large claims all claims above the 99th quantile of the original distribution of costs. Together, these claims represented almost 7% of all costs. Even after the exclusion of large claims, the distribution was still skewed and heavy-tailed, which seriously complicated the choice of a distribution to be used in GLM. We concluded that the distributions that best fit the data were Lognormal, Pareto and Gamma distributions, in this order. However, the fact that the first two distributions do not belong to the exponential family left us no alternative besides returning to the Gamma distribution.

Besides using multiple imputation and choosing wisely which distribution must be used in each model, we explored Generalized Additive Models (GAM) as an alternative to GLM. The advantage of using GAM instead of GLM is having the opportunity to apply smooth functions to covariates, making them able to capture data tendencies different from linear. Also, because of its capacity of smoothing cycling variables, smooth functions are an excellent tool to adjust models for seasonal effects.

One of the hardest tasks developed in this dissertation was combining the methods explored before. Even though most methods are implemented in R, the packages which implement them are very specific and usually developed only for applying a certain method. As a consequence, it is common that the packages are incompatible. Therefore, so we could combine the methods previously explored into a single model, we defined a methodology which one should follow in order to achieve the frequency and severity models defined for the final tariff.

In the end, we concluded that, despite the classical approach being very reasonable, the final tariff seems to add quality to the models, improving their ability to predict new data. More specifically, we conclude that GAM seem to produce models more capable of predicting test data sets than GLM and that multiple imputation is an

excellent method to work around the possible bias caused by complete cases analysis, since a tariff constructed based on a data set where a set of observations were excluded may not correctly estimate the reality, especially if MAR is assumed.

6.2 Future Work

Defining and developing a methodology to create a tariff involves consecutive advances and setbacks in many steps along the way, which most times means making several decisions so that the work can proceed. For that reason, there are several branches in this tree that have been cut, each one indicating which paths to be explored.

The most evident path is related to the assumptions made in section 1.2.1. In fact, it is easy to think in hypothetical situations that would make us believe that these assumptions are not real, in particular the time independence assumption. Therefore, studying the validity of this assumption is a possible work topic to be considered. Even more, in case any of the assumptions are proven to be unrealistic, we lose the property of all observations in our data set being independent, which might mean that different types of models must be assumed, as for example mixed effects models.

"Essentially, all models are wrong, but some are useful."

George Box, 1978

References

H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, Dec. 1974.

M. Azur, E. Stuart, C. Frangakis, and P. Leaf. Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 3 2011. ISSN 1049-8931. doi: 10.1002/mpr.329.

R. E. Beard, T. Pentikinen, and E. Pesonen. *Risk Theory: The Stochastic Basis of Insurance*. Monographs on Statistics and Applied Probability 20. Springer Netherlands, 1 edition, 1984. ISBN 9789401176828.

J. Beirlant and Y. Goegebeur. Regression with response distributions of pareto-type. *Computational Statistics & Data Analysis*, 42(4):595 – 619, 2003. ISSN 0167-9473.

E. Boelviken. *Computation and Modelling in Insurance and Finance*. CUP, 2014. ISBN 9780521830485.

C. D. Boor and G. H. Golub. *Recent Advances in Numerical Analysis. Proceedings of a Symposium Conducted by the Mathematics Research Center, the University of Wisconsin-Madison*. Publication of the Mathematics Research Center, the University of Wisconsin-Madison. Elsevier Inc, Academic Press Inc, 1978. ISBN 9780122083600.

A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data*. CUP, 2ed. edition, 2013. ISBN 9781107014169.

R. R. Corbeil and S. R. Searle. Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38, 1976.

M. L. Delignette-Muller and C. Dutang. fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34, 2015. URL http://www.jstatsoft.org/v64/i04/.

A. J. Dobson. *An Introduction to Generalized Linear Models*. Chapman & Hall, 2 edition, 2001. ISBN 1584881658.

J. Duchon. *Splines minimizing rotation-invariant semi-norms in Sobolev spaces*, pages 85–100. Springer Berlin Heidelberg, Berlin, Heidelberg, 1977. ISBN 9783540374961.

O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. ISSN 01621459. URL http://www.jstor.org/stable/2282330.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics. Springer-Verlag New York, 2 edition, 2001. ISBN 9781441929006.

J. W. Graham, A. E. Olchowski, and T. D. Gilreath. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3):206–213, 2007. doi: 10.1007/s11121-007-0070-9.

P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Monographs on Statistics and Applied Probability 58. Springer US, 1994. ISBN 9780412300400.

C. Gu. *Smoothing Spline ANOVA Models*. Springer Series in Statistics 297. Springer-Verlag New York, 2 edition, 2013. ISBN 9781461453680.

T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, 1 edition, 1990. ISBN 9780412343902.

J. M. Hilbe. *Negative Binomial Regression, Second Edition*. Cambridge University Press, 2 edition, 2011. ISBN 0521198151,9780521198158.

J. M. Hilbe. *Modeling Count Data*. Cambridge University Press, 1 edition, 2014. ISBN 9781107611252.

V. Hoef, J. M., Boveng, and P. L. Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772, November 2007. ISSN 0012-9658.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615733.

S. Jackman. *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory*. United States Studies Centre, University of Sydney, Sydney, New South Wales, Australia, 2017. URL https://github.com/atahk/pscl/. R package version 1.5.1.

R. Kaas, M. Goovaerts, J. Dhaene, and M. Denuit. *Modern Actuarial Risk Theory, Using R*. Springer, 2nd ed.2008. corr. 2nd printing edition, 2009. ISBN 9783540709923.

L. J. Keele. *Semiparametric Regression for the Social Sciences*. 1 edition, 2008. ISBN 9780470319918.

C. Kleiber and A. Zeileis. *Applied Econometrics with R*. Springer-Verlag, New York, 2008. URL https://CRAN.R-project.org/package=AER. ISBN 9780387773162.

C. Lanczos. An iterative method for the solution of the eigenvalue problem of linear differential and integral, 1950.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238.

E. Ohlsson and B. Johansson. *Non-life insurance pricing with generalized linear models*. EAA lecture notes. Springer-Verlag Berlin Heidelberg, 1 edition, 2010. ISBN 9783642107900.

P. Parodi. *Pricing in General Insurance*. CRC Press, 2014. ISBN 9781466581449.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

Z. Reitermanov. Data splitting. *Proceedings of the 19th Annual Conference of Doctoral Students*, pages 31–36, 2010.

D. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library. Wiley, 1987. ISBN 9780471655749.

J. L. Schafer and M. K. Olsen. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research*, 33:545–571, 1998.

G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978.

J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338, 2009. ISSN 0959-8138.

S. K. Thompson. *Sampling*. Wiley Series in Probability and Statistics. Wiley, 3 edition, 2012. ISBN 9780470402313.

M. A. A. Turkman and G. L. Silva. *Modelos Lineares Generalizados, da teoria prtica*. Edies SPE, Lisboa, 2000.

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. URL http://www.jstatsoft.org/v45/i03/.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, pages 307–333, 1989.

P. Wilson. The misuse of the vuong test for non-nested models to test for zero-inflation. *Economics Letters*, 127:51 – 53, 2015. ISSN 0165-1765.

S. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2006. ISBN 9781584884743.

S. Wood, N., Pya, and B. S"afken. Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*, 111:1548–1575, 2016.

S. N. Wood. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, 65(1):95–114, 2003.

S. N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99 (467):673–686, 2004.

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

A. Zeileis, C. Kleiber, and S. Jackman. Regression models for count data in R. *Journal of Statistical Software*, 27(8), 2008. URL http://www.jstatsoft.org/v27/i08/.

A. F. Zuur, E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. *Mixed Effects Models and Extensions in Ecology with R*. Statistics for Biology and Health. Springer-Verlag New York, 1 edition, 2009. ISBN 0387874577.